SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Mislav Zorko

UTVRĐIVANJE INDIKATORA USPJEŠNOSTI STUDIRANJA

Diplomski rad

Voditelj rada:
prof. dr. sc. Miljenko Marušić

Zagreb, rujan, 2015

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

U ovom diplomskom radu promatramo uzorak upisanih studenata na studij matematike u razdoblju od 2005. godine do 2011. godine. Za svakog studenta iz baze fakulteta promatramo godinu upisa na studij, broj postignutih bodova u srednjoj školi, broj postignutih bodova na odgovarajućem prijemnom ispitu, te položenost kolegija kroz prvu godinu studiranja. Studenti slušaju ukupno osam kolegija kroz prvu godinu studiranja:

1. Prvi semestar
    a) Matematička analiza 1 (oznaka: MA1, broj ECTS bodova: 8)
    b) Linearna algebra 1 (oznaka: LA1, broj ECTS bodova: 8)
    c) Elementarna matematika 1 (oznaka: EM1, broj ECTS bodova: 8)
    d) Programiranje 1 (oznaka: PR1, broj ECTS bodova: 6)
2. Drugi semestar
    a) Matematička analiza 2 (oznaka: MA2, broj ECTS bodova: 9)
    b) Linearna algebra 2 (oznaka: LA2, broj ECTS bodova: 9)
    c) Elementarna matematika 2 (oznaka: EM2, broj ECTS bodova: 6)
    d) Programiranje 2 (oznaka: PR2, broj ECTS bodova: 6)

Definirat ćemo pojam uspješnog studenta kao studenta koji je u roku položio sve kolegije prve godine. Također ćemo zasebno promatrati i studente koji su položili sve kolegije prvog semestra u roku. Takve modele zbog preglednosti označavamo simbolom ($\triangle$).

Prvo poglavlje rada odnosi se na deskriptivnu statistiku podataka. U prvom djelu dajemo opće oznake i promatramo utjecaj promjene računarskih kolegija kroz odgovarajuće godine. Analizirati ćemo posebno bodove iz srednje škole, a zatim i bodove postignute na odgovarajućim prijemnim ispitima. Također promatramo zasebno uspješne i neuspješne studente. Na kraju prvog poglavlja analiziramo uspješnost studiranja po promatranim kolegijima i godinama.

Drugo i treće poglavlje bave se teorijskom razradom logističke regresije. Promatramo univarijatni i multivarijatni model. Definirani su opći pojmovi koje koristimo kroz rad, obrađena je teorijska pozadina prilagodbe modela logističke regresije. Obrađeni su testovi značajnosti koeficijenata modela i pripadni intervali pouzdanosti.

Četvrto poglavlje bavi se primjenom logističke regresije na promatrani uzorak. Zasebno promatramo dio uzorka prije uvođenja državne mature i poslije, te dio uzorka sa studentima koji su uspješno položili sve kolegije iz prvog semestra (modeli $\triangle$). Dajemo modele univarijatne logističke regresije za nezavisne varijable $\mathcal{BS}$ (bodovi iz srednje škole) i $\mathcal{BT}$ (bodovi na odgovarajućem prijemnom ispitu). Također su dana i četiri modela multivarijatne logističke regresije, dva za promatrani uzorak prije uvođenja državne mature i dva za promatrani uzorak nakon uvođenja. U tablici 0.1 dajemo pregled razrađenih modela.

<table>
<tr><th></th><th>Univarijatni model</th><th>Multivarijatni model</th></tr>
<tr><td rowspan="4">2005. - 2009.</td><td>$\mathcal{BS}$</td><td rowspan="2">$\mathcal{BS}$ + $\mathcal{BT}$</td></tr>
<tr><td>$\mathcal{BT}$</td></tr>
<tr><td>$\mathcal{BS}$ ($\triangle$)</td><td rowspan="2">$\mathcal{BS}$ ($\triangle$) + $\mathcal{BT}$ ($\triangle$)</td></tr>
<tr><td>$\mathcal{BT}$ ($\triangle$)</td></tr>
<tr><td rowspan="4">2010. - 2011.</td><td>$\mathcal{BS}$</td><td rowspan="2">$\mathcal{BS}$ + $\mathcal{BT}$</td></tr>
<tr><td>$\mathcal{BT}$</td></tr>
<tr><td>$\mathcal{BS}$ ($\triangle$)</td><td rowspan="2">$\mathcal{BS}$ ($\triangle$) + $\mathcal{BT}$ ($\triangle$)</td></tr>
<tr><td>$\mathcal{BT}$ ($\triangle$)</td></tr>
</table>

Tablica 0.1: Pregled modela

U dodatku rada dajemo algoritme za procjenu parametara modela logističke regresije pomoću iterativne težinske metode najmanjih kvadrata. Općeniti kod za multivarijatni (ali i univarijatni) model implementiran je u programskom jeziku Matlab-u.

Rad je u cijelosti pisan u programu za obradu teksta LaTeX, dok je statistički dio napravljen pomoću programskih jezika za statistiku R i SAS.

Poglavlje 1

Deskriptivna analiza studenata na smjeru matematika

1.1 Uvod i opće oznake

Uzorak na kojem provodimo analizu čine svi studenti koji su upisali studij matematike 2005., 2006., 2007., 2008., 2009., 2010. i 2011. godine. Za svakog upisanog studenta promatramo sljedeće varijable:

i) godinu upisa

ii) broj bodova postignutih u srednjoj školi (oznaka: $\mathcal{BS}$)

iii) broj bodova postignutih na odgovarajućem prijemnom ispitu (oznaka: $\mathcal{BT}$)

iv) dihotomne varijable uspjeha po kolegijima

Promatramo prvo računarske kolegije koji se pojavljuju u danim podacima. Koristimo se sljedećim oznakama:

- UUP - Uvod u programiranje
- PROG (C) - Programiranje (C)
- PROG 1 - Programiranje 1
- PROG 2 - Programiranje 2

Tablica 1.1 prikazuje razliku u računarskim kolegijima kroz promatrane godine, te raspon bodova iz srednje škole i raspon bodova na prijemnom ispitu.

Godina	Predmet	Predmet	raspon $\mathcal{BS}$	raspon $\mathcal{BT}$
2005	UUP	PROG (C)	0-260	0-663
2006	UUP	PROG (C)	0-260	0-663
2007	PROG 1	PROG 2	0-260	0-663
2008	PROG 1	PROG 2	0-260	0-663
2009	PROG 1	PROG 2	0-260	0-663
2010	PROG 1	PROG 2	0-300	0-600
2011	PROG 1	PROG 2	0-300	0-600

Tablica 1.1: Prikaz računarskih kolegija i upisnih bodova po godinama

Kako postoji promjena između računarskih kolegija, u samom uvodnom djelu provodimo usporedbu između kolegija Uvod u programiranje i Programiranje 1, te između Programiranje (C) i Programiranje 2. Neka je $G = \{2005, 2006, 2007, 2008, 2009, 2010, 2011\}$ skup godina koje promatramo i neka su $a, b \in G$. Za $i \in G$ stavimo da je n_i veličina uzorka upisana u i-toj godini, a X_i pripadni dio uzorka koji je položio pripadni računarski kolegij. Testiramo hipotezu o jednakosti parametara za godine a i b za računarske kolegije u zimskom semestru (Uvod u programiranje i Programiranje 1), odnosno testiramo

$$H_0 : p_a = p_b. \tag{1.1}$$

Definiramo

$$p_{a,b} := \frac{X_a + X_b}{n_a + n_b} \, (= p_{b,a}), \tag{1.2}$$

te računamo testnu statistiku

$$Z_{a,b} = \frac{\hat{p}_a - \hat{p}_b}{\sqrt{p_{a,b}(1 - p_{a,b})}\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}} \overset{H_0}{\sim} AN(0, 1) \tag{1.3}$$

U tablici 1.2 nalaze se pripadne p-vrijednosti testova između godina.

g	2005	2006	2007	2008	2009	2010	2011
2005	1.00000	0.16087	0.89848	0.07389	0.15652	0.00029	0.00000
2006	0.16087	1.00000	0.14907	0.64644	0.00592	0.00000	0.00000
2007	0.89848	0.14907	1.00000	0.07053	0.22408	0.00089	0.00000
2008	0.07389	0.64644	0.07053	1.00000	0.00221	0.00000	0.00000
2009	0.15652	0.00592	0.22408	0.00221	1.00000	0.03074	0.00036
2010	0.00029	0.00000	0.00089	0.00000	0.03074	1.00000	0.15086
2011	0.00000	0.00000	0.00000	0.00000	0.00036	0.15086	1.00000

Tablica 1.2: Rezultati testa poroporcija za računarske kolegije zimskog semestra

Iz tablice 1.2 vidimo da za $\alpha = 0.05$ postoji statistički značajna razlika između nekih godina, ali ne možemo primijetiti utjecaj promjene kolegija. Isti test provodimo i za računarske kolegije ljetnog semestra (Programiranje (C) i Programiranje 2), te dobivamo pripadne p- vrijednosti koje dajemo u tablici 1.3

g	2005	2006	2007	2008	2009	2010	2011
2005	1.00000	0.00298	0.19116	0.00197	0.01060	0.86048	0.21810
2006	0.00298	1.00000	0.13431	0.76889	0.75258	0.00233	0.09981
2007	0.19116	0.13431	1.00000	0.08890	0.24855	0.15201	0.91680
2008	0.00197	0.76889	0.08890	1.00000	0.55899	0.00154	0.06514
2009	0.01060	0.75258	0.24855	0.55899	1.00000	0.00820	0.19755
2010	0.86048	0.00233	0.15201	0.00154	0.00820	1.00000	0.17338
2011	0.21810	0.09981	0.91680	0.06514	0.19755	0.17338	1.00000

Tablica 1.3: Testa poroporcija za računarske kolegije ljetnog semestra

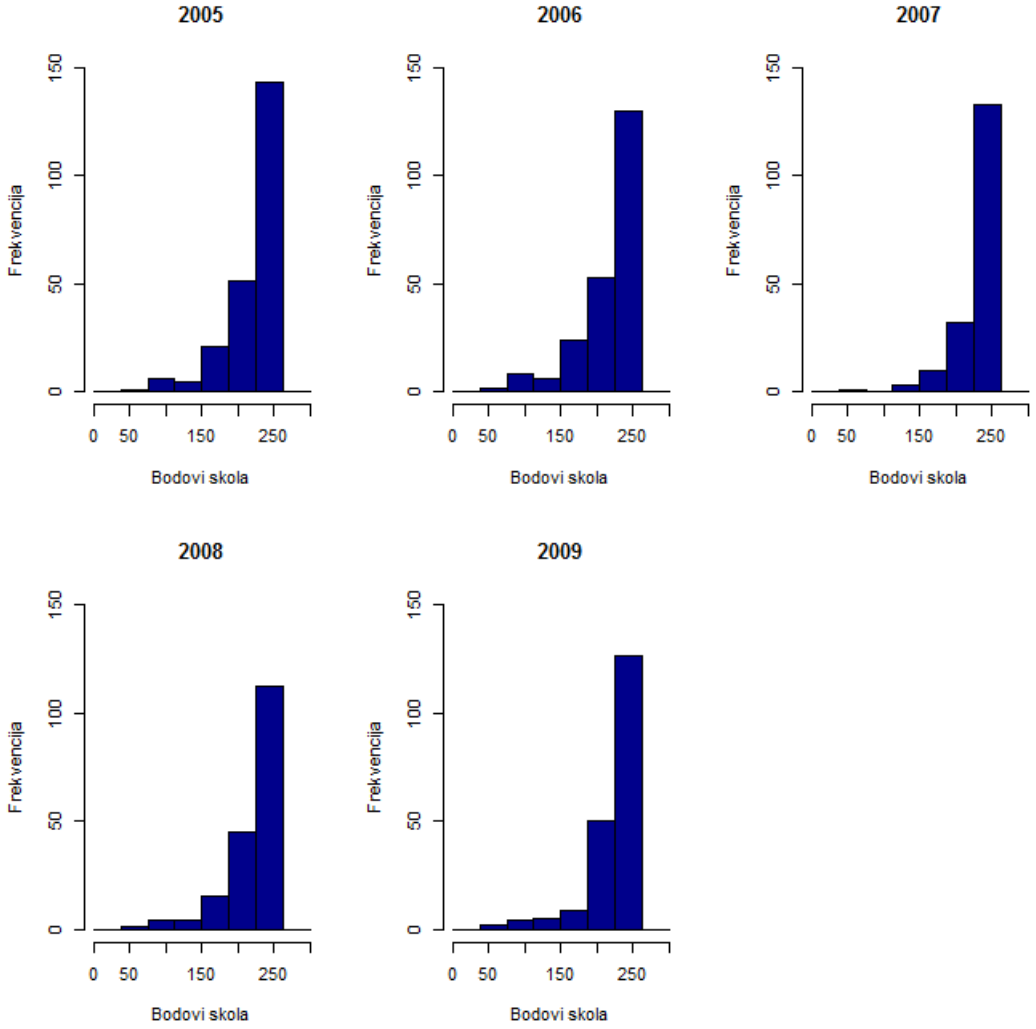
Na temelju navedenoga zaključujemo da promjena kolegija ne utječe značajno na prolaznost, te zbog jednostavnosti možemo pretpostaviti da studenti slušaju sljedeće kolegije:

1. Prvi semestar
    a) Matematička analiza 1 (oznaka: MA1, broj ECTS bodova: 8)
    b) Linearna algebra 1 (oznaka: LA1, broj ECTS bodova: 8)
    c) Elementarna matematika 1 (oznaka: EM1, broj ECTS bodova: 8)
    d) Programiranje 1 (oznaka: PR1, broj ECTS bodova: 6)
2. Drugi semestar
    a) Matematička analiza 2 (oznaka: MA2, broj ECTS bodova: 9)
    b) Linearna algebra 2 (oznaka: LA2, broj ECTS bodova: 9)
    c) Elementarna matematika 2 (oznaka: EM2, broj ECTS bodova: 6)
    d) Programiranje 2 (oznaka: PR2, broj ECTS bodova: 6)

Kako je raspon bodova i način bodovanja prije uvođenja državne mature drugačiji nego nakon uvođenja državne mature, radit ćemo nezavisne analize. Prvo ćemo analizirati bodove iz srednje škole, zatim bodove prijemnog ispita/državne mature, te na kraju analizirati uspješnost studenata.

1.2 Analiza bodova iz srednje škole

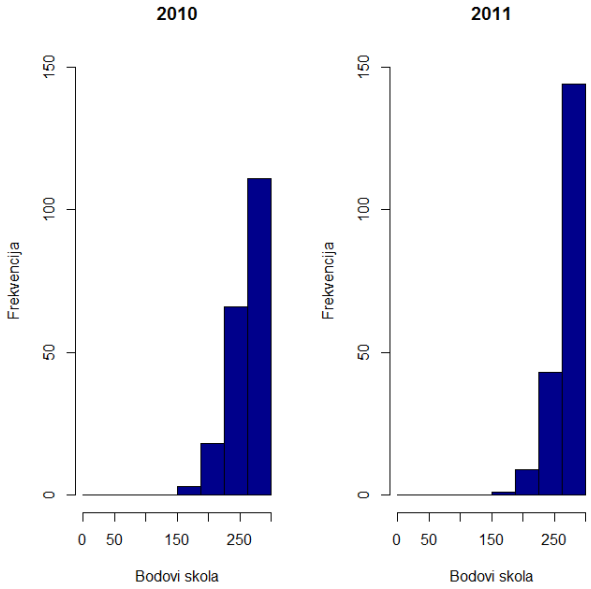
Iz tablice 1.1 vidimo da je od 2005. do uključeno 2009. godine maksimalan broj bodova iz srednje škole bio 260. Od 2010. godine, zbog državne mature, maksimalan broj bodova postaje 300. Prvo promatramo bodove iz srednje škole za studente upisane od 2005. do 2009. godine uključeno. Dobivamo sljedeće stupčaste dijagrame prikazane na slici 1.1.



Slika 1.1: Stupčasti dijagrami bodova iz srednje škole (2005.-2009.)

Sa slike 1.1 vidljivo je da preko pola upisanih studenata ima vrlo visok broj bodova i uočavamo rast frekvencija po broju bodova. Na slici 1.2 prikazani su stupčasti dijagrami

frekvencija bodova iz srednje škole za studente upisane 2010. i 2011. godine.



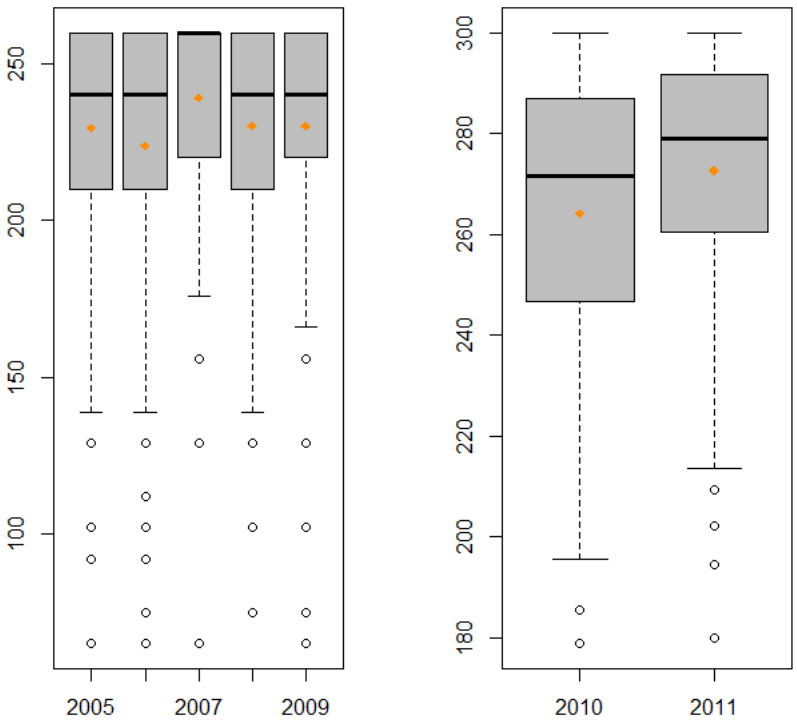
Slika 1.2: Stupčasti dijagrami bodova iz srednje škole (2010., 2011.)

U tablici 1.4 se nalaze osnovne statističke veličine za bodove iz srednje škole po godinama.

	2005	2006	2007	2008	2009		2010	2011
Minimum	65.0	65.0	65.0	75.0	65.0	Minimum	178.8	180.0
Prvi kvartil	210.0	210.0	220.0	210.0	220.0	Prvi kvartil	246.9	260.4
Medijan	240.0	240.0	260.0	240.0	240.0	Medijan	271.5	279.0
Srednja vr.	229.5	223.9	239.0	230.1	230.1	Srednja vr.	264.1	272.6
Treći kvartil	260.0	260.0	260.0	260.0	260.0	Treći kvartil	286.7	291.6
Maksimum	260.0	260.0	260.0	260.0	260.0	Maksimum	300.0	300.0

Tablica 1.4: Osnovne statističke veličine bodova iz škole po godinama

Također, dobivamo i sljedeće box-plotove prikazane na slici 1.3. Sve godine stavljene su zajedno, bez obzira na razliku u rasponu bodova.



Slika 1.3: Box-plot po godinama za bodove iz srednje škole

Promotrimo detaljnije uzoračko očekivanje (oznaka $\mu_g,\ g \in G$) i standardnu devijaciju (oznaka $s_g,\ g \in G$) danih uzoraka. Podaci su dani u tablici 1.5

Provodimo test analize varijance (ANOVA), odnosno testiramo hipotezu

$$H_0 : \mu_{2005} = \mu_{2006} = \mu_{2007} = \mu_{2008} = \mu_{2009}. \tag{1.4}$$

Dobivene rezultate dajemo u tablici 1.6.

g	n_g	μ_g	s_g	s_g^2
2005	227	229.5	39.558	1564.817
2006	223	223.9	43.855	1923.248
2007	179	239.0	31.433	988.0498
2008	181	230.1	38.362	1471.66
2009	196	230.1	38.835	1508.154
2010	198	264.1	28.444	809.062
2011	197	272.6	23.888	570.642

Tablica 1.5: Uzoračko očekivanje i standardna devijacija

	Stupnjevi slobode	Suma kvadrata	Kvadrat očekivanja	F statistika	p-vrijednost
Godina	4	22638	5659	3.738	0.00501
Reziduali	1001	1515471	1514	-	-

Tablica 1.6: ANOVA test za bodove iz škole

Iz tablice 1.6 vidljivo je da na nivou značajnosti $\alpha = 0.05$ postoji statistički značajna razlika između barem dvije grupe. Provodimo Tukey test i dobivamo rezultate dane u tablici 1.7.

godina	razlika srednjih vrijednosti	95% interval pouzdanosti	p-vrijednost
2006-2005	−5.55176705	$\langle -15.577093, 4.473559 \rangle$	0.5539923
2007-2005	9.48308025	$\langle -1.145645, 20.111806 \rangle$	0.1061158
2008-2005	0.64360990	$\langle -9.952232, 11.239452 \rangle$	0.9998299
2009-2005	0.61305403	$\langle -9.754751, 10.980860 \rangle$	0.9998471
2007-2006	15.03484731	$\langle 4.364177, 25.705518 \rangle$	0.0011827
2008-2006	6.19537695	$\langle -4.442540, 16.833294 \rangle$	0.5031594
2009-2006	6.16482109	$\langle -4.245980, 16.575623 \rangle$	0.4859847
2008-2007	−8.83947035	$\langle -20.047856, 2.368915 \rangle$	0.1978538
2009-2007	−8.87002622	$\langle -19.863089, 2.123036 \rangle$	0.1786053
2009-2008	−0.03055587	$\langle -10.991828, 10.930717 \rangle$	1.0000000

Tablica 1.7: Tukey test razlike uzoračkih očekivanja grupa

Iz tablice 1.7 vidimo da postoji statistički značajna razlike između 2007. i 2006. godine, što se moglo predvidjeti iz grafičkog prikaza 1.3.

Testiramo još i jednakost uzoračkih očekivanja $\mu_{2010} = \mu_{2011}$, odnosno testiramo hipotezu

$$H_0 : \mu_{2010} = \mu_{2011}. \tag{1.5}$$

Testna statistika koju koristimo je

$$Z_{a,b} = \frac{\mu_a - \mu_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}, \tag{1.6}$$

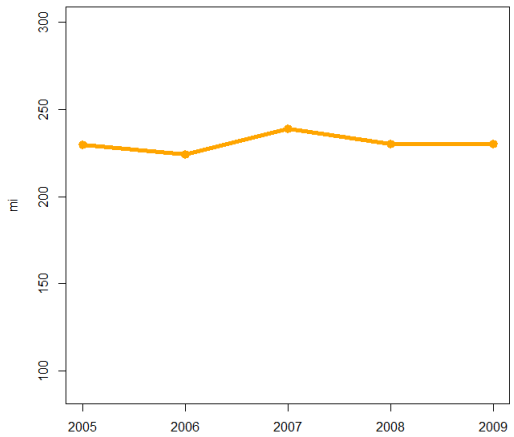
te pod pretpostavkom da vrijedi hipoteza H_0 imamo $Z \overset{H_0}{\sim} AN(0,1)$, odnosno uz nultu hipotezu o jednakostima očekivanja imamo da $Z \xrightarrow{\mathcal{D}} N(0,1)$ kada $n_a,\ n_b \longrightarrow +\infty$. Prema tome, za dovoljno velike n_g računamo pripadnu p-vrijednost kao $2 \cdot (\mathbb{P}(|Z| > z)) = 2 \cdot (1 - \mathbb{P}(|Z| \leq z))$.

Dobivena p- vrijednost za testnu hipotezu

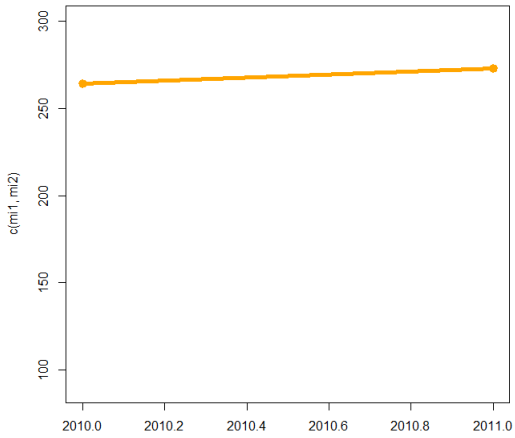
$$H_0 : \mu_{2010} = \mu_{2011} \tag{1.7}$$

je $p = 0.001254343$, dakle postoji i statistički značajna razlika između μ_{2010} i μ_{2011} (na nivou značajnosti $\alpha = 0.05$).

Na slici 1.4 i 1.5 dan je grafički prikaz srednjih vrijednosti bodova iz škole za upisane studente.

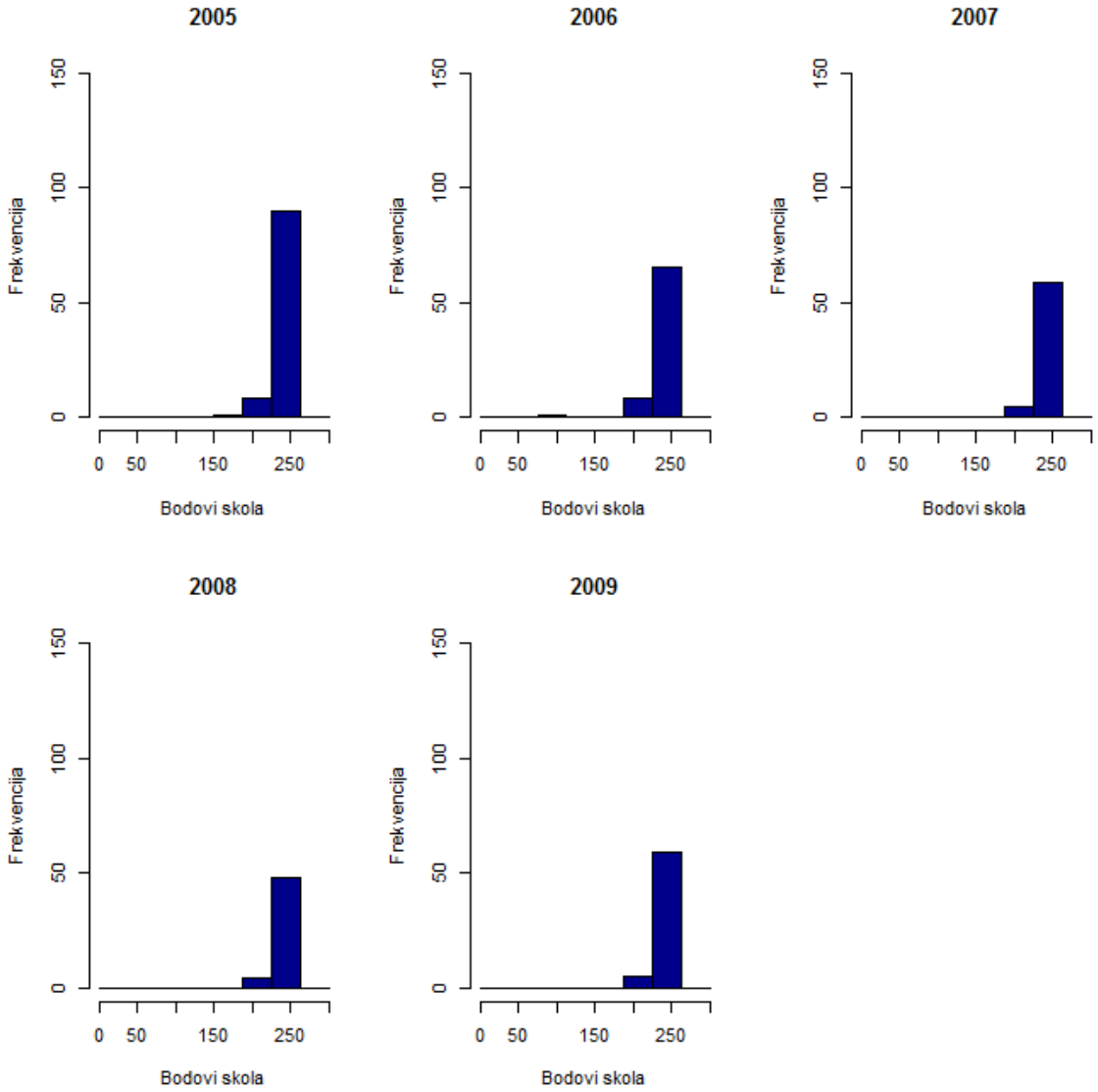


Slika 1.4: Srednje vrijednosti bodova iz škole

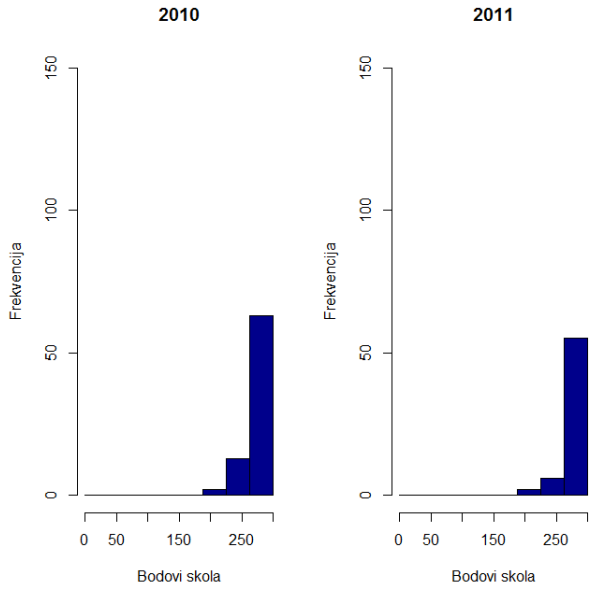


Slika 1.5: Srednje vrijednosti bodova iz škole

Promotrimo sada bodove iz srednje škole samo za uspješne studente. Dobivamo sljedeće stupčaste dijagrame prikazane na slici 1.6 i 1.7. Sa slika je vidljivo da uspješni studenti imaju u pravilu više bodova iz srednje škole.



Slika 1.6: Stupčasti dijagrami za uspješne studente (2005.-2009.)



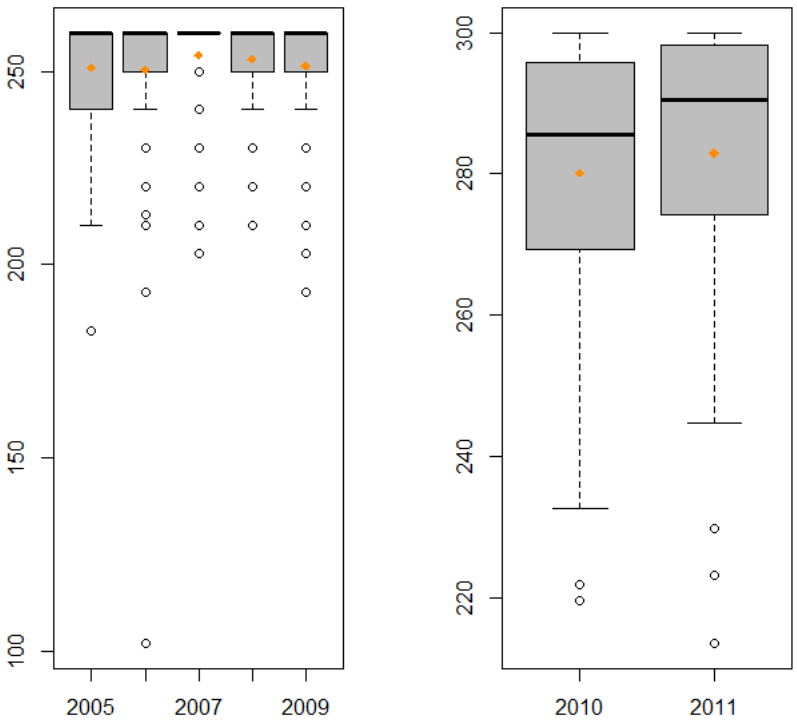
Slika 1.7: Stupčasti dijagrami za uspješne studente (2010., 2011.)

U tablici 1.8 se nalaze osnovne statističke veličine za bodove iz srednje škole po godinama za uspješne studente.

	2005	2006	2007	2008	2009		2010	2011
Minimum	183.0	102.0	203.0	210.0	193.0	Minimum	219.6	213.6
Prvi kvartil	240.0	250.0	260.0	250.0	250.0	Prvi kvartil	270.0	274.2
Medijan	260.0	260.0	260.0	260.0	260.0	Medijan	285.6	290.4
Srednja vr.	250.9	250.4	254.1	253.1	251.3	Srednja vr.	280.1	282.9
Treći kvartil	260.0	260.0	260.0	260.0	260.0	Treći kvartil	295.8	298.1
Maksimum	260.0	260.0	260.0	260.0	260.0	Maksimum	300.0	300.0

Tablica 1.8: Osnovne statističke veličine bodova iz škole po godinama za uspješne studente

Također, dobivamo i sljedeće box-plotove prikazane na slici 1.8. Sve godine stavljene su zajedno, bez obzira na razliku u rasponu bodova.



Slika 1.8: Box-plot za bodove iz srednje škole

Promotrimo detaljnije uzoračko očekivanje (oznaka $\mu_g,\ g \in G$) i standardnu devijaciju (oznaka $s_g,\ g \in G$) danih uzoraka. Podaci su dani u tablici 1.9

Provodimo test analize varijance (ANOVA), odnosno testiramo hipotezu

$$H_0 : \mu_{2005} = \mu_{2006} = \mu_{2007} = \mu_{2008} = \mu_{2009} \tag{1.8}$$

samo za uspješne studente.

Dobivene rezultate dajemo u tablici 1.10.

g	n_g	μ_g	s_g	s_g^2
2005	99	250.9	15.844	251.037
2006	74	250.4	23.235	539.882
2007	64	254.1	13.468	181.401
2008	52	253.1	13.216	174.661
2009	64	251.3	15.554	241.944
2010	78	280.1	19.844	393.786
2011	63	282.6	19.376	375.444

Tablica 1.9: Uzoračko očekivanje i standardna devijacija za uspješne studente

	Stupnjevi slobode	Suma kvadrata	Kvadrat očekivanja	F statistika	p-vrijednost
Godina	4	662	165.6	0.579	0.678
Residuali	348	99591	286.2	-	-

Tablica 1.10: ANOVA test za bodove iz škole za uspješne studente

Iz tablice 1.10 vidljivo je da ne postoji statistički značajna razlika između grupa, odnosno da su sva uzoračka očekivanja jednaka.

Testiramo još i jednakost uzoračkih očekivanja $\mu_{2010} = \mu_{2011}$, odnosno testiramo hipotezu

$$H_0 : \mu_{2010} = \mu_{2011}. \tag{1.9}$$

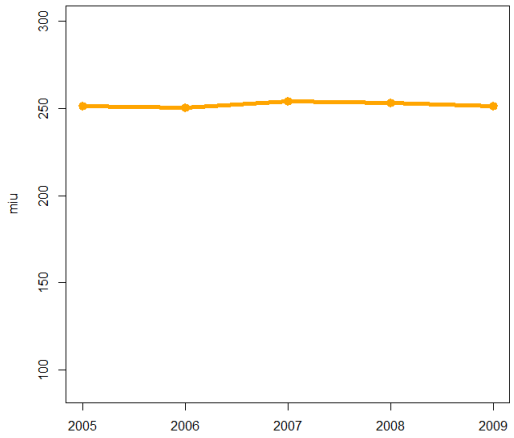
Testna statistika koju koristimo je

$$Z_{a,b} = \frac{\mu_a - \mu_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}, \tag{1.10}$$

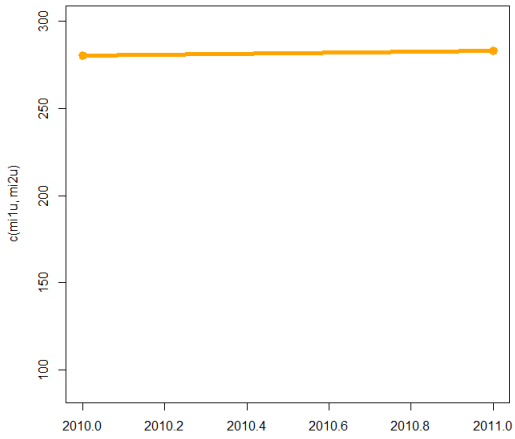
te pod pretpostavkom da vrijedi hipoteza H_0 imamo $Z \overset{H_0}{\sim} AN(0,1)$, odnosno uz nultu hipotezu o jednakostima očekivanja imamo da $Z \xrightarrow{\mathcal{D}} N(0,1)$ kada $n_a,\ n_b \longrightarrow +\infty$. Prema tome, za dovoljno velike n_g računamo pripadnu p-vrijednost kao $2 \cdot (\mathbb{P}(|Z| > z)) = 2 \cdot (1 - \mathbb{P}(|Z| \leq z))$. Dobivena p- vrijednost za testnu hipotezu

$$H_0 : \mu_{2010} = \mu_{2011} \tag{1.11}$$

je $p = 0.406$, dakle ne postoji statistički značajna razlika između μ_{2010} i μ_{2011} (na nivou značajnosti $\alpha = 0.05$). Na slici 1.9 i 1.10 dan je grafički prikaz srednjih vrijednosti bodova iz škole za uspješne studente.

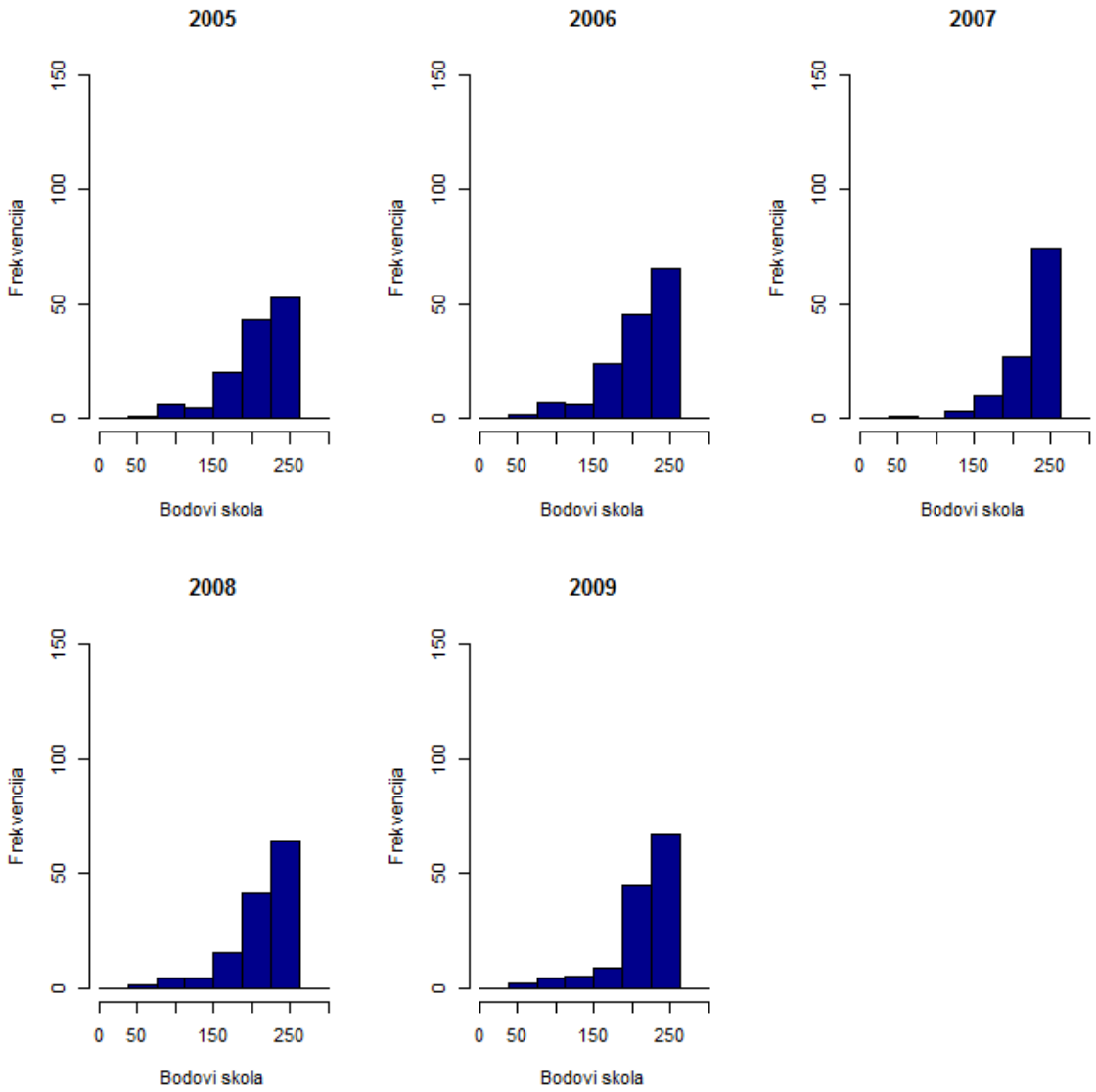


Slika 1.9: Srednje vrijednosti bodova iz škole za uspješne studente

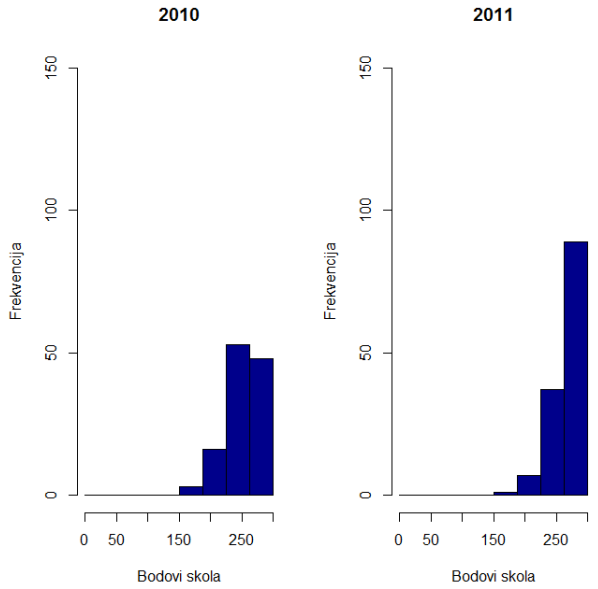


Slika 1.10: Srednje vrijednosti bodova iz škole za uspješne studente

Na kraju ovog poglavlja, promotrimo bodove iz srednje škole samo za neuspješne studente. Dobivamo sljedeće stupčaste dijagrame prikazane na slici 1.11 i 1.12. Sa slika je vidljivo da neuspješni studenti imaju u pravilu manje bodova iz srednje škole nego uspješni studenti. Također je vidljiv usporeni rast frekvencija po broju bodova.



Slika 1.11: Stupčasti dijagrami za neuspješne studente (2005.-2009.)



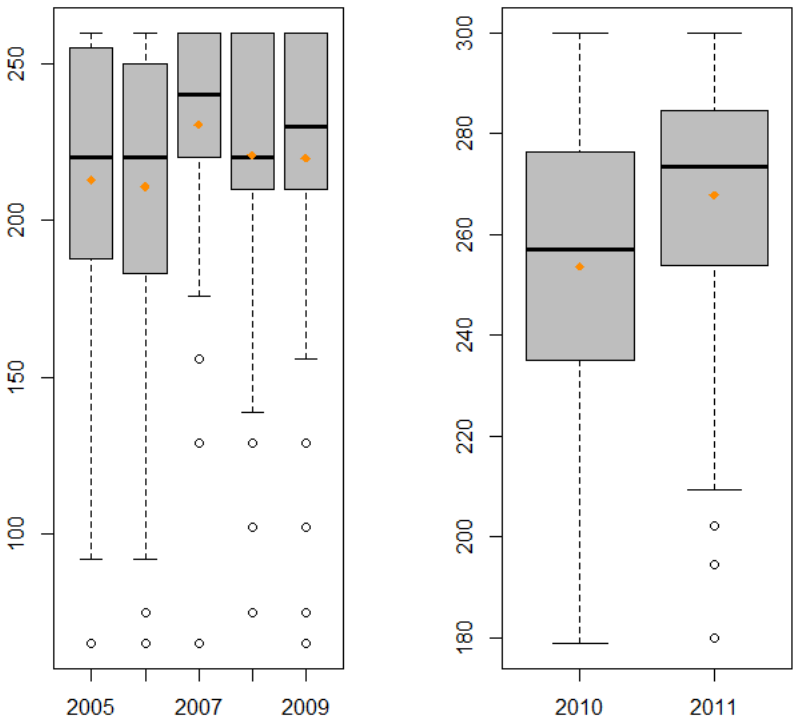
Slika 1.12: Stupčasti dijagrami za neuspješne studente (2010., 2011.)

U tablici 1.11 se nalaze osnovne statističke veličine za bodove iz srednje škole po godinama za neuspješne studente.

	2005	2006	2007	2008	2009		2010	2011
Minimum	65.0	65.0	65.0	75.0	65.0	Minimum	178.9	180.0
Prvi kvartil	190.5	183.0	220.0	210.0	210.0	Prvi kvartil	235.0	254.0
Medijan	220.0	220.0	240.0	220.0	230.0	Medijan	257.1	254.0
Srednja vr.	212.9	210.8	230.5	220.9	219.8	Srednja vr.	253.6	267.8
Treći kvartil	252.5	250.0	260.0	260.0	260.0	Treći kvartil	275.8	294.2
Maksimum	260.0	260.0	260.0	260.0	260.0	Maksimum	300.0	300.0

Tablica 1.11: Osnovne statističke veličine bodova iz škole po godinama za neuspješne studente

Također, dobivamo i sljedeće box-plotove prikazane na slici 1.13. Sve godine stavljene su zajedno, bez obzira na razliku u rasponu bodova.



Slika 1.13: Box-plot za bodove iz srednje škole

Promotrimo detaljnije uzoračko očekivanje (oznaka $\mu_g,\ g \in G$) i standardnu devijaciju (oznaka $s_g,\ g \in G$) danih uzoraka. Podaci su dani u tablici 1.12

Provodimo test analize varijance (ANOVA), odnosno testiramo hipotezu

$$H_0 : \mu_{2005} = \mu_{2006} = \mu_{2007} = \mu_{2008} = \mu_{2009} \tag{1.12}$$

samo za neuspješne studente.

Dobivene rezultate dajemo u tablici 1.13.

g	n_g	μ_g	s_g	s_g^2
2005	128	212.9	44.213	1954.832
2006	149	210.8	45.776	2095.401
2007	115	230.5	35.246	1242.267
2008	129	220.9	41.229	1699.854
2009	132	219.8	42.441	1801.289
2010	120	253.6	28.385	805.6815
2011	134	267.8	24.334	592.1693

Tablica 1.12: Uzoračko očekivanje i standardna devijacija za neuspješne studente

	Stupnjevi slobode	Suma kvadrata	Kvadrat očekivanja	F statistika	p-vrijednost
Godina	4	30488	7622	4.282	0.00199
Residuali	648	1153552	1780	-	-

Tablica 1.13: ANOVA test za bodove iz škole za neuspješne studente

Iz tablice 1.13 vidljivo je da na nivou značajnosti $\alpha = 0.05$ postoji statistički značajna razlika između barem dvije grupe. Provodimo Tukey test i dobivamo rezultate dane u tablici 1.14.

godina	razlika srednjih vrijednosti	95% interval pouzdanosti	p-vrijednost
2006-2005	−2.093068	$\langle -16.002372, 11.816235\rangle$	0.9939712
2007-2005	17.649389	$\langle 2.820355, 32.478422\rangle$	0.0104252
2008-2005	7.985283	$\langle -6.413650, 22.384217\rangle$	0.5517258
2009-2005	6.904593	$\langle -7.412615, 21.221801\rangle$	0.6793443
2007-2006	19.742457	$\langle 5.416504, 34.068410\rangle$	0.0016674
2008-2006	10.078352	$\langle -3.801922, 23.958625\rangle$	0.2738497
2009-2006	8.997661	$\langle -4.797814, 22.793137\rangle$	0.3837132
2008-2007	−9.664105	$\langle -24.465913, 5.137702\rangle$	0.3826138
2009-2007	−10.744796	$\langle -25.467114, 3.977523\rangle$	0.2688337
2009-2008	−1.080691	$\langle -15.369698, 13.208316\rangle$	0.9995919

Tablica 1.14: Tukey test razlike uzoračkih očekivanja grupa

Iz tablice 1.14 vidimo da postoji statistički značajna razlika između 2007. i 2006. godine, te između 2007. i 2005. godine što se moglo predvidjeti iz grafičkog prikaza 1.13.

Testiramo još i jednakost uzoračkih očekivanja $\mu_{2010} = \mu_{2011}$, odnosno testiramo hipotezu

$$H_0 : \mu_{2010} = \mu_{2011}. \tag{1.13}$$

Testna statistika koju koristimo je

$$Z_{a,b} = \frac{\mu_a - \mu_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}, \tag{1.14}$$

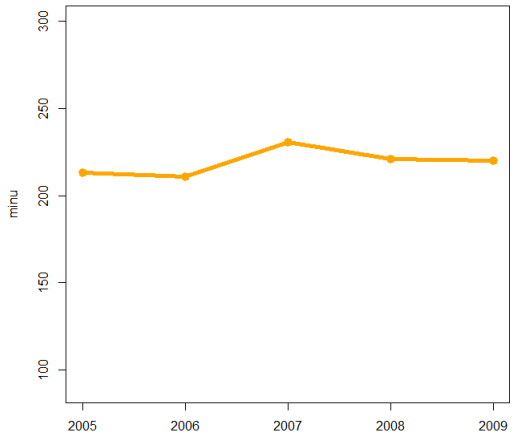
te pod pretpostavkom da vrijedi hipoteza H_0 imamo $Z \overset{H_0}{\sim} AN(0,1)$, odnosno uz nultu hipotezu o jednakostima očekivanja imamo da $Z \xrightarrow{\mathcal{D}} N(0,1)$ kada $n_a,\ n_b \longrightarrow +\infty$. Prema tome, za dovoljno velike n_g računamo pripadnu p-vrijednost kao $2 \cdot (\mathbb{P}(|Z| > z)) = 2 \cdot (1 - \mathbb{P}(|Z| \leq z))$.

Dobivena p- vrijednost za testnu hipotezu

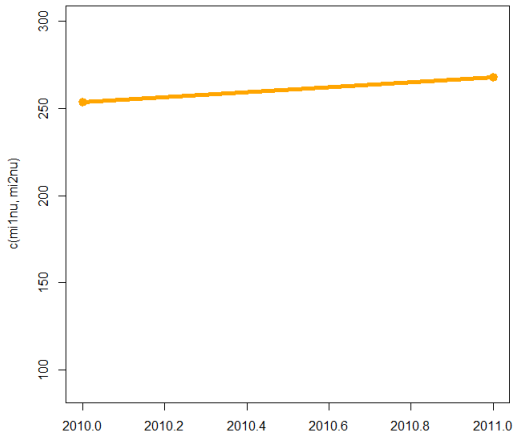
$$H_0 : \mu_{2010} = \mu_{2011} \tag{1.15}$$

je izrazito mala, dakle postoji statistički značajna razlika između μ_{2010} i μ_{2011} (na nivou značajnosti $\alpha = 0.05$).

Na slici 1.14 i 1.15 dan je grafički prikaz srednjih vrijednosti bodova iz škole za neuspješne studente.



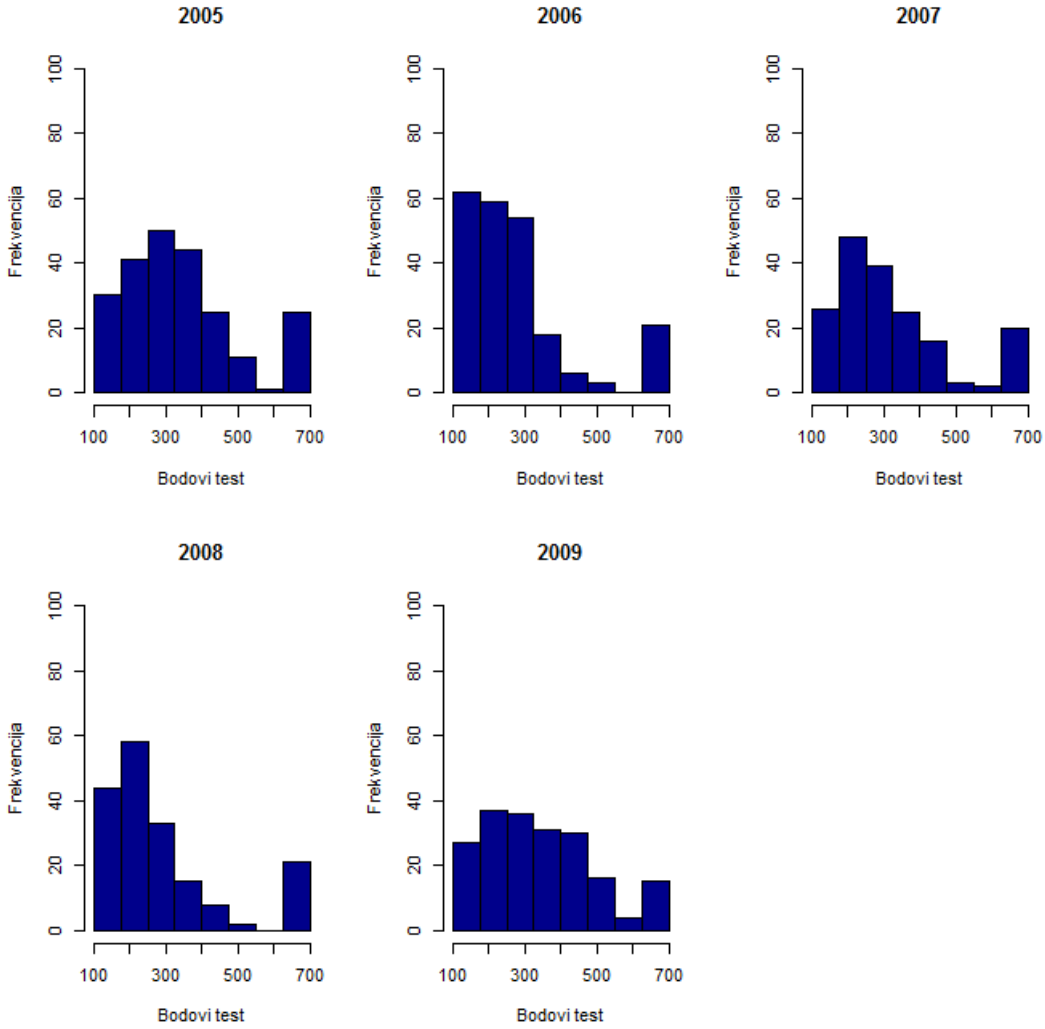
Slika 1.14: Srednje vrijednosti bodova iz škole za neuspješne studente



Slika 1.15: Srednje vrijednosti bodova iz škole za neuspješne studente

1.3 Analiza bodova s prijemnog ispita/državne mature

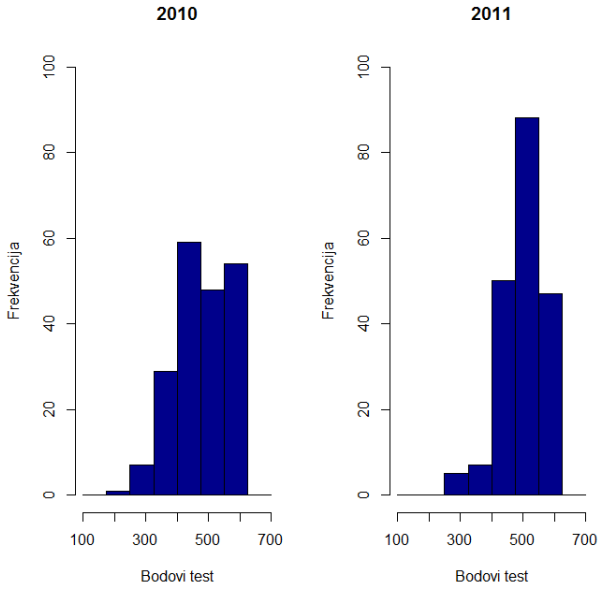
Iz tablice 1.1 vidimo da je od 2005. do uključeno 2009. godine maksimalan broj bodova na prijemnom ispitu bio 663. Od 2010. godine, zbog državne mature, maksimalan broj bodova postaje 600. Prvo promatramo bodove prijemnog ispita za studente upisane od 2005. do 2009. godine uključeno. Dobivamo sljedeće stupčaste dijagrame prikazane na slici 1.16.



Slika 1.16: Stupčasti dijagrami za bodove prijemnog ispita (2005.-2009.)

Sa slike 1.16 vidljivo je da ... Na slici 1.17 prikazani su stupčasti dijagrami frekvencija

bodova državne mature za studente upisane 2010. i 2011. godine.



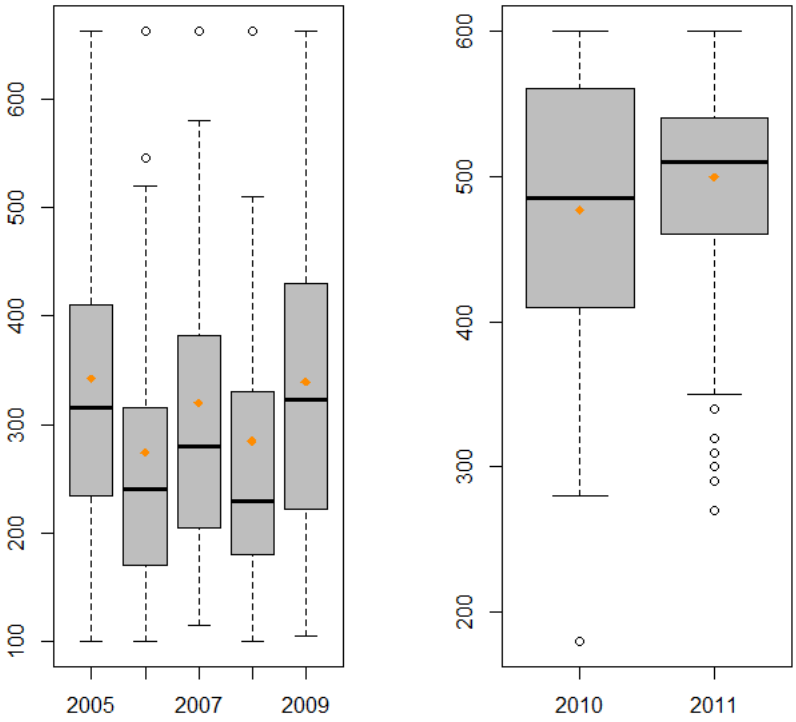
Slika 1.17: Stupčasti dijagrami za bodove iz srednje škole (2010., 2011.)

U tablici 1.15 se nalaze osnovne statističke veličine za prijemnog ispita, odnosno državne mature po godinama.

	2005	2006	2007	2008	2009		2010	2011
Minimum	100.0	100.0	115.0	100.0	105.0	Minimum	180.0	270.0
Prvi kvartil	235.0	170.0	205.0	180.0	223.0	Prvi kvartil	410.0	460.0
Medijan	315.0	240.0	280.0	230.0	322.5	Medijan	485.0	510.0
Srednja vr.	342.1	274.3	320.3	284.9	339.2	Srednja vr.	476.6	499.4
Treći kvartil	410.0	315.0	382.5	330.0	430.0	Treći kvartil	560.0	540.0
Maksimum	663.0	663.0	663.0	663.0	663.0	Maksimum	600.0	600.0

Tablica 1.15: Osnovne statističke veličine bodova prijemnog ispita

Također, dobivamo i sljedeće box-plotove prikazane na slici 1.18. Sve godine stavljene su zajedno, bez obzira na razliku u rasponu bodova.



Slika 1.18: Box-plot za bodove prijemnog ispita/državne mature

Promotrimo detaljnije uzoračko očekivanje (oznaka $\mu_g,\ g \in G$) i standardnu devijaciju (oznaka $s_g,\ g \in G$) danih uzoraka. Podaci su dani u tablici 1.16

g	n_g	μ_g	s_g	s_g^2
2005	227	342.1	150.710	22713.54
2006	223	274.3	152.036	23114.85
2007	179	320.3	155.036	24039.66
2008	181	284.9	162.677	26463.81
2009	196	339.2	148.374	22014.77
2010	198	476.6	87.423	7642.768
2011	197	499.4	63.248	4000.246

Tablica 1.16: Uzoračko očekivanje i standardna devijacija

Provodimo test analize varijance (ANOVA), odnosno testiramo hipotezu

$$H_0 : \mu_{2005} = \mu_{2006} = \mu_{2007} = \mu_{2008} = \mu_{2009}. \tag{1.16}$$

Dobivene rezultate dajemo u tablici 1.17.

	Stupnjevi slobode	Suma kvadrata	Kvadrat očekivanja	F statistika	p-vrijednost
Godina	4	813470	203367	8.626	7.61e-07
Residuali	1001	23600182	23577	-	-

Tablica 1.17: ANOVA test za bodove prijemnog ispita

Iz tablice 1.17 vidljivo je da na nivou značajnosti $\alpha = 0.05$ postoji statistički značajna razlika između barem dvije grupe. Provodimo Tukey test i dobivamo rezultate dane u tablici 1.18.

godina	razlika srednjih vrijednosti	95% interval pouzdanosti	p-vrijednost
2006-2005	−67.826910	$\langle -107.389281, -28.264539 \rangle$	0.0000313
2007-2005	−21.814584	$\langle -63.758115, 20.128947 \rangle$	0.6140204
2008-2005	−57.298951	$\langle -99.112717, -15.485186 \rangle$	0.0017830
2009-2005	−2.966106	$\langle -43.879983, 37.947771 \rangle$	0.9996566
2007-2006	46.012326	$\langle 3.903271, 88.121380 \rangle$	0.0241302
2008-2006	10.527959	$\langle -31.451842, 52.507760 \rangle$	0.9596565
2009-2006	64.860804	$\langle 23.777254, 105.944353 \rangle$	0.0001708
2008-2007	−35.484367	$\langle -79.715377, 8.746644 \rangle$	0.1833583
2009-2007	18.848478	$\langle -24.532815, 62.229771 \rangle$	0.7587075
2009-2008	54.332845	$\langle 11.077003, 97.588686 \rangle$	0.0056111

Tablica 1.18: Tukey test razlike uzoračkih očekivanja grupa

Iz tablice 1.18 vidimo da postoji statistički značajna razlika (na nivou značajnosti $\alpha = 0.05$) između 2006. i 2005., 2008. i 2005., 2007. i 2006., 2009. i 2006. godine, te između 2009. i 2008. godine što se moglo predvidjeti iz grafičkog prikaza 1.18.

Testiramo još i jednakost uzoračkih očekivanja $\mu_{2010} = \mu_{2011}$, odnosno testiramo hipotezu

$$H_0 : \mu_{2010} = \mu_{2011}. \tag{1.17}$$

Testna statistika koju koristimo je

$$Z_{a,b} = \frac{\mu_a - \mu_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}, \tag{1.18}$$

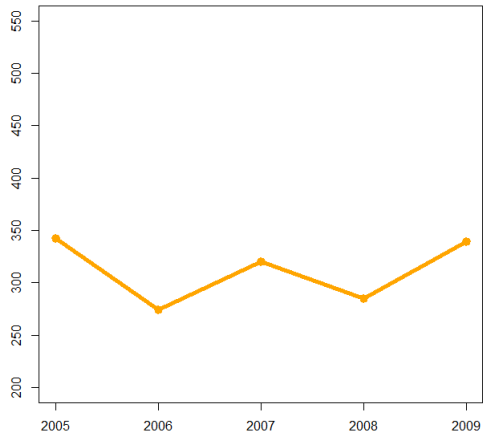
te pod pretpostavkom da vrijedi hipoteza H_0 imamo $Z \overset{H_0}{\sim} AN(0,1)$, odnosno uz nultu hipotezu o jednakostima očekivanja imamo da $Z \xrightarrow{\mathcal{D}} N(0,1)$ kada $n_a,\ n_b \longrightarrow +\infty$. Prema tome, za dovoljno velike n_g računamo pripadnu p-vrijednost kao $2 \cdot (\mathbb{P}(|Z| > z)) = 2 \cdot (1 - \mathbb{P}(|Z| \leq z))$.

Dobivena p-vrijednost za testnu hipotezu

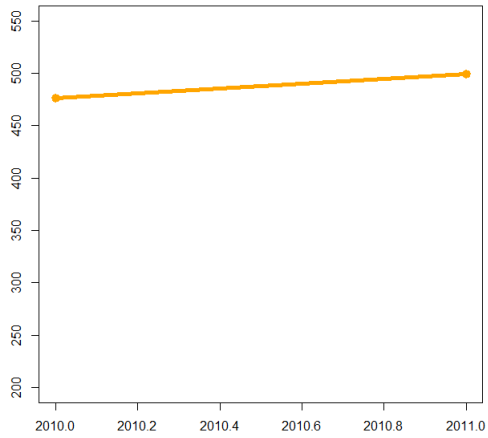
$$H_0 : \mu_{2010} = \mu_{2011} \tag{1.19}$$

je $p = 0.002969925$, dakle postoji i statistički značajna razlika između μ_{2010} i μ_{2011} (na nivou značajnosti $\alpha = 0.05$).

Na slici 1.19 i 1.20 dan je grafički prikaz srednjih vrijednosti bodova prijemnog ispita, odnosno državne mature za upisane studente.

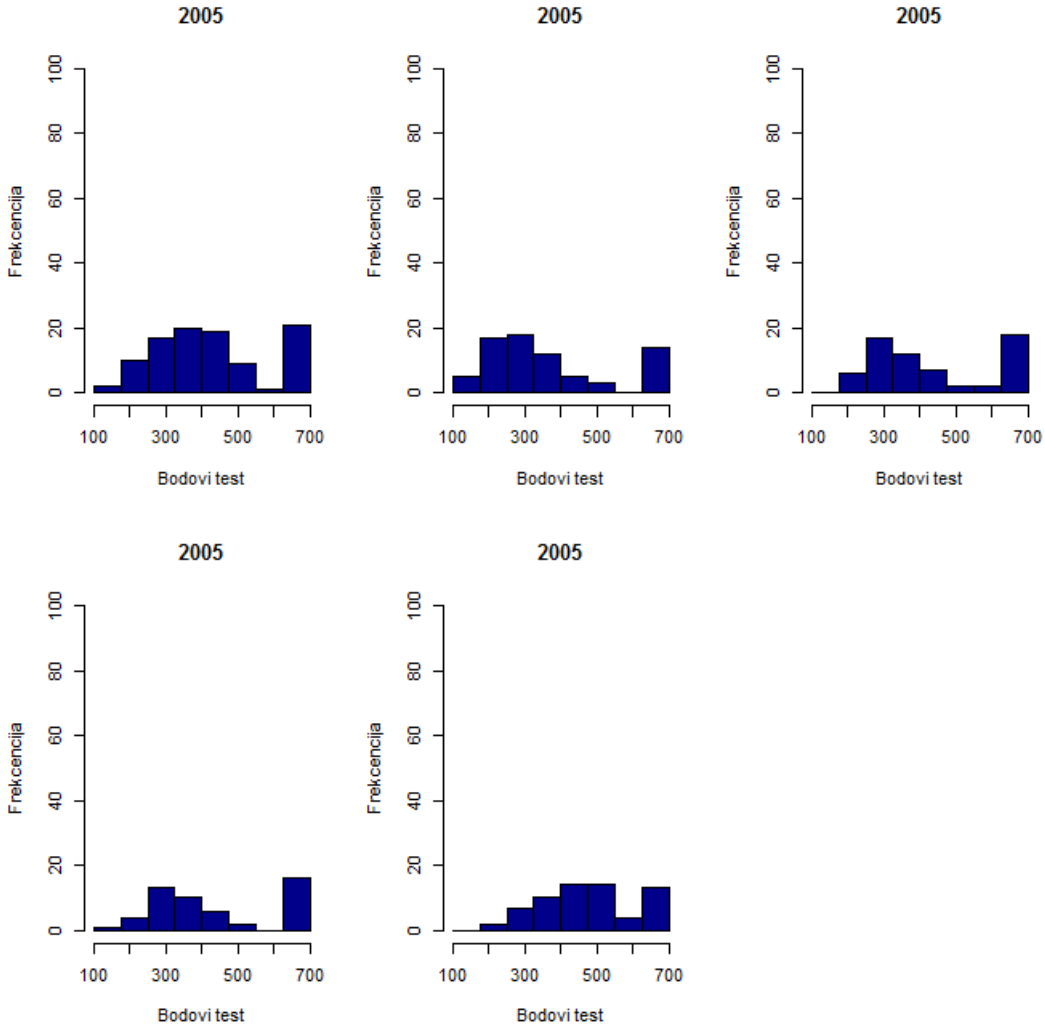


Slika 1.19: Srednje vrijednosti bodova prijemnog za sve studente

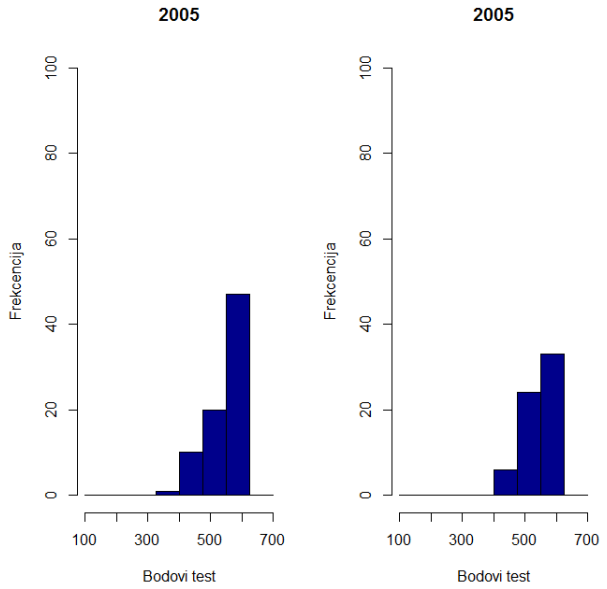


Slika 1.20: Srednje vrijednosti bodova prijemnog za sve studente

Promotrimo sada bodove prijemnog ispita, odnosno državne mature samo za uspješne studente. Dobivamo sljedeće stupčaste dijagrame prikazane na slici 1.21 i 1.22. Sa slika je vidljivo da uspješni studenti imaju u pravilu više bodova iz srednje škole.



Slika 1.21: Stupčasti dijagrami za uspješne studente (2005.-2009.)



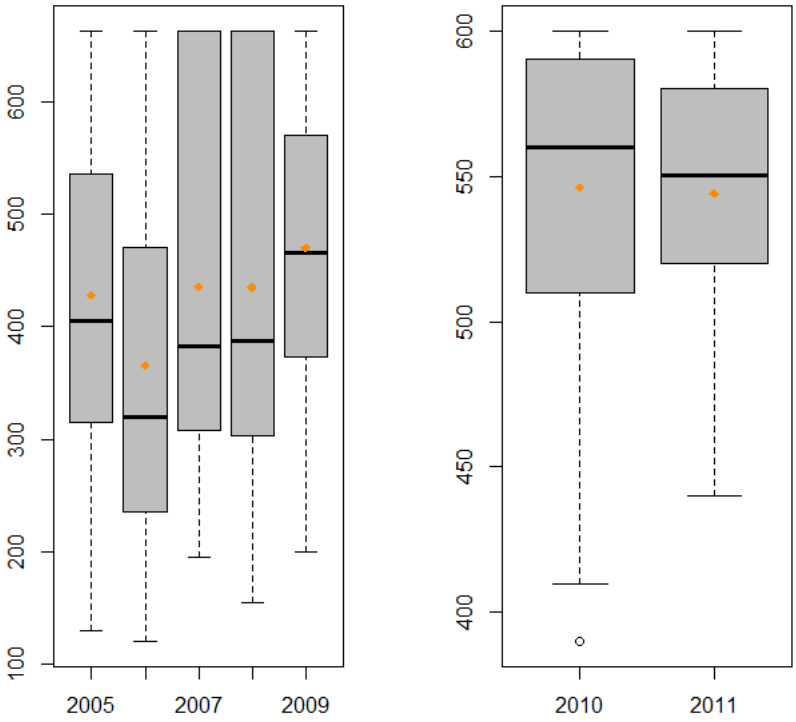
Slika 1.22: Stupčasti dijagrami za uspješne studente (2010., 2011.)

U tablici 1.19 se nalaze osnovne statističke veličine za bodove iz srednje škole po godinama za uspješne studente.

	2005	2006	2007	2008	2009		2010	2011
Minimum	130.0	120.0	195.0	155.0	200.0	Minimum	390.0	440.0
Prvi kvartil	315.0	237.5	308.8	306.2	376.2	Prvi kvartil	512.5	520.0
Medijan	405.0	320.0	382.5	387.5	465.0	Medijan	560.0	550.0
Srednja vr.	428.0	364.8	435.7	434.7	469.8	Srednja vr.	546.0	544.0
Treći kvartil	535.0	466.2	663.0	663.0	570.0	Treći kvartil	590.0	580.0
Maksimum	663.0	663.0	663.0	663.0	663.0	Maksimum	600.0	600.0

Tablica 1.19: Osnovne statističke veličine bodova prijemnog ispita/državne mature po godinama za uspješne studente

Također, dobivamo i sljedeće box-plotove prikazane na slici 1.23. Sve godine stavljene su zajedno, bez obzira na razliku u rasponu bodova.



Slika 1.23: Box-plot za bodove prijemnog ispita/državne mature

Promotrimo detaljnije uzoračko očekivanje (oznaka $\mu_g,\ g \in G$) i standardnu devijaciju (oznaka $s_g,\ g \in G$) danih uzoraka. Podaci su dani u tablici 1.20

g	n_g	μ_g	s_g	s_g^2
2005	99	428.0	149.903	22471.00
2006	74	364.8	169.335	28674.00
2007	64	435.7	162.790	26500.5
2008	52	434.7	169.426	28705.17
2009	64	469.8	131.288	17236.63
2010	78	546.0	52.272	2732.329
2011	63	544.0	39.904	1592.327

Tablica 1.20: Uzoračko očekivanje i standardna devijacija za uspješne studente

Provodimo test analize varijance (ANOVA), odnosno testiramo hipotezu

$$H_0 : \mu_{2005} = \mu_{2006} = \mu_{2007} = \mu_{2008} = \mu_{2009}. \tag{1.20}$$

Dobivene rezultate dajemo u tablici 1.21.

	Stupnjevi slobode	Suma kvadrata	Kvadrat očekivanja	F statistika	p-vrijednost
Godina	4	409282	102320	4.182	0.00254
Residuali	348	8514795	24468	-	-

Tablica 1.21: ANOVA test za bodove prijemnog ispita

Iz tablice 1.21 vidljivo je da na nivou značajnosti $\alpha = 0.05$ postoji statistički značajna razlika između barem dvije grupe. Provodimo Tukey test i dobivamo rezultate dane u tablici 1.22.

godina	razlika srednjih vrijednosti	95% interval pouzdanosti	p-vrijednost
2006-2005	−63.189189	$\langle -129.102909, 2.72453 \rangle$	0.0674011
2007-2005	7.687500	$\langle -61.109966, 76.48497 \rangle$	0.9980782
2008-2005	6.673077	$\langle -66.787685, 80.13384 \rangle$	0.9991480
2009-2005	41.750000	$\langle -27.047466, 110.54747 \rangle$	0.4576536
2007-2006	70.876689	$\langle -2.341658, 144.09504 \rangle$	0.0630776
2008-2006	69.862266	$\langle -7.754218, 147.47875 \rangle$	0.1003359
2009-2006	104.939189	$\langle 31.720842, 178.15754 \rangle$	0.0009669
2008-2007	−1.014423	$\langle -81.094330, 79.06548 \rangle$	0.9999997
2009-2007	34.062500	$\langle -41.762293, 109.88729 \rangle$	0.7327617
2009-2008	35.076923	$\langle -45.002984, 115.15683 \rangle$	0.7507578

Tablica 1.22: Tukey test razlike uzoračkih očekivanja grupa

Iz tablice 1.22 vidimo da postoji statistički značajna razlika (na nivou značajnosti $\alpha = 0.05$) između 2009. i 2006. godine što se moglo predvidjeti iz grafičkog prikaza 1.23.

Testiramo još i jednakost uzoračkih očekivanja $\mu_{2010} = \mu_{2011}$, odnosno testiramo hipotezu

$$H_0 : \mu_{2010} = \mu_{2011}. \tag{1.21}$$

Testna statistika koju koristimo je

$$Z_{a,b} = \frac{\mu_a - \mu_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}, \tag{1.22}$$

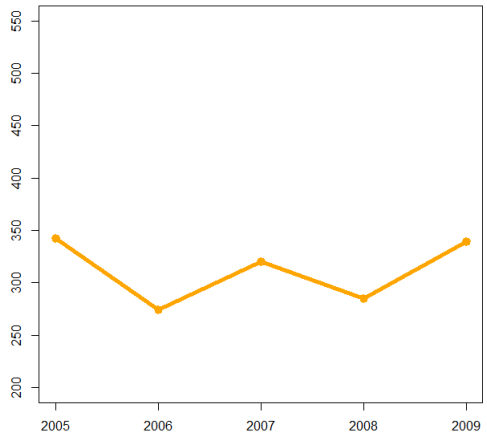
te pod pretpostavkom da vrijedi hipoteza H_0 imamo $Z \overset{H_0}{\sim} AN(0,1)$, odnosno uz nultu hipotezu o jednakostima očekivanja imamo da $Z \xrightarrow{\mathcal{D}} N(0,1)$ kada $n_a,\ n_b \longrightarrow +\infty$. Prema tome, za dovoljno velike n_g računamo pripadnu p-vrijednost kao $2 \cdot (\mathbb{P}(|Z| > z)) = 2 \cdot (1 - \mathbb{P}(|Z| \leq z))$.

Dobivena p-vrijednost za testnu hipotezu

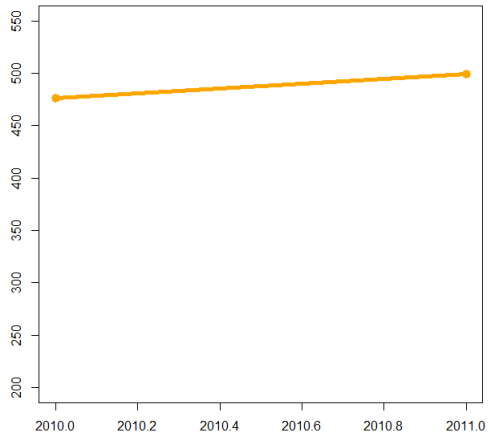
$$H_0 : \mu_{2010} = \mu_{2011} \tag{1.23}$$

je $p = 0.7978976$, dakle ne postoji statistički značajna razlika između μ_{2010} i μ_{2011} (na nivou značajnosti $\alpha = 0.05$).

Na slici 1.24 i 1.25 dan je grafički prikaz srednjih vrijednosti bodova prijemnog ispita, odnosno državne mature za uspješne studente.



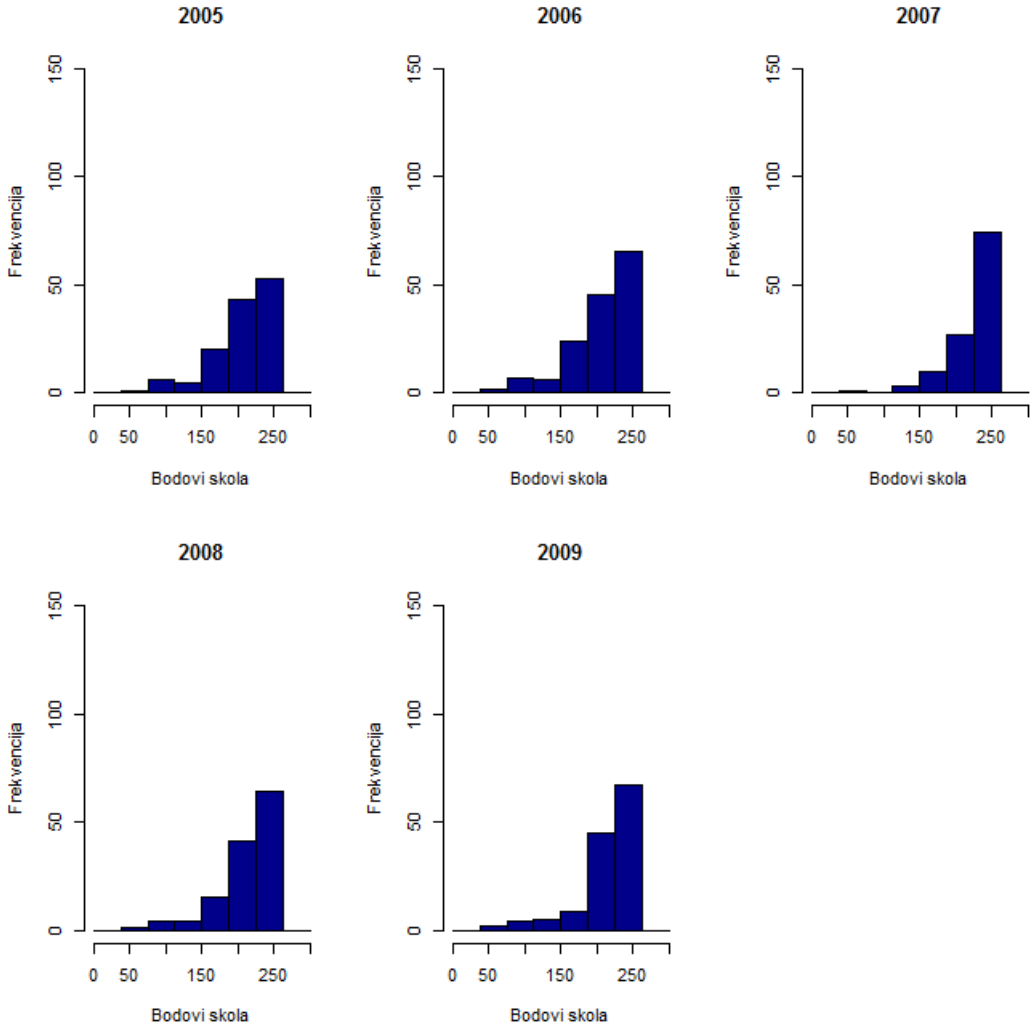
Slika 1.24: Srednje vrijednosti bodova prijemnog za uspješne studente



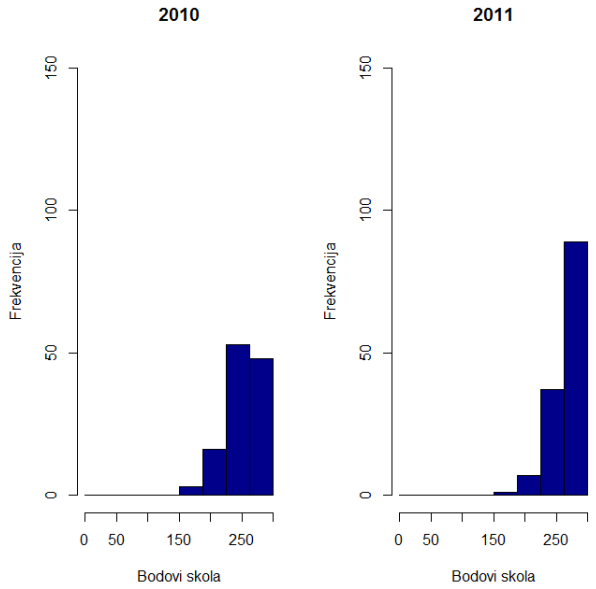
Slika 1.25: Srednje vrijednosti bodova prijemnog za uspješne studente

Usporedimo li slike 1.19 i 1.20 sa slikama 1.24 i 1.25 vidljivo je da uspješni studenti imaju u prosjeku više bodova ostvarenih na prijemnom (državnoj maturi) prosjeka ostvarenih bodova svih studenata.

Na kraju ovog poglavlja promotrimo bodove prijemnog ispita (državne mature) samo za neuspješne studente. Dobivamo sljedeće stupčaste dijagrame prikazane na slici 1.26 i 1.27. Sa slika je vidljivo da neuspješni studenti imaju u pravilu manje bodova iz srednje škole nego uspješni studenti. Također je vidljiv usporen rast frekvencija po broju bodova.



Slika 1.26: Stupčasti dijagrami za neuspješne studente (2005.-2009.)



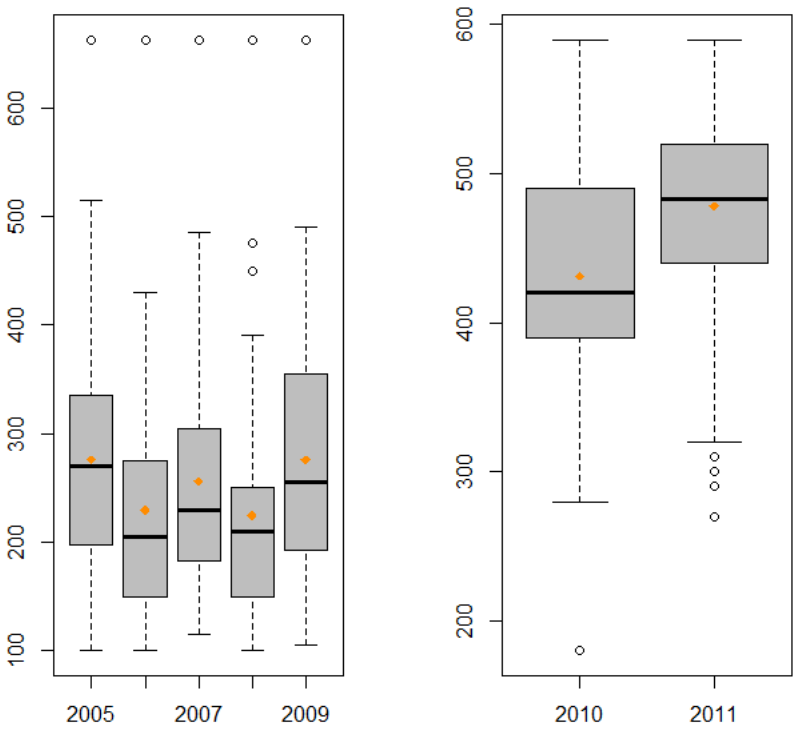
Slika 1.27: Stupčasti dijagrami za neuspješne studente (2010., 2011.)

U tablici 1.23 se nalaze osnovne statističke veličine za bodove prijemnog ispita/državne mature po godinama za neuspješne studente.

	2005	2006	2007	2008	2009		2010	2011
Minimum	100.0	100.0	115.0	100.0	105.0	Minimum	180.0	270.0
Prvi kvartil	198.8	150.0	182.5	150.0	193.8	Prvi kvartil	390.0	442.5
Medijan	270.0	205.0	230.0	210.0	255.0	Medijan	420.0	482.8
Srednja vr.	275.8	229.4	256.1	224.5	275.9	Srednja vr.	431.5	478.4
Treći kvartil	335.0	275.0	305.0	250.0	355.0	Treći kvartil	490.0	520.0
Maksimum	663.0	663.0	663.0	663.0	663.0	Maksimum	590.0	590.0

Tablica 1.23: Osnovne statističke veličine bodova prijemnog ispita/državne mature po godinama za neuspješne studente

Također, dobivamo i sljedeće box-plotove prikazane na slici 1.28. Sve godine stavljene su zajedno, bez obzira na razliku u rasponu bodova.



Slika 1.28: Box-plot za bodove prijemnog ispita i državne mature

Promotrimo detaljnije uzoračko očekivanje (oznaka $\mu_g,\ g \in G$) i standardnu devijaciju (oznaka $s_g,\ g \in G$) danih uzoraka. Podaci su dani u tablici 1.24

g	n_g	μ_g	s_g	s_g^2
2005	128	275.8	113.537	12890.58
2006	149	229.4	120.006	14401.49
2007	115	256.1	106.128	11263.19
2008	129	224.5	113.942	12982.67
2009	132	275.9	110.064	12114.02
2010	120	431.5	75.347	5677.095
2011	134	478.4	61.381	3767.657

Tablica 1.24: Uzoračko očekivanje i standardna devijacija za neuspješne studente

Provodimo test analize varijance (ANOVA), odnosno testiramo hipotezu

$$H_0 : \mu_{2005} = \mu_{2006} = \mu_{2007} = \mu_{2008} = \mu_{2009} \tag{1.24}$$

za neuspješne studente. Dobivene rezultate dajemo u tablici 1.25.

	Stupnjevi slobode	Suma kvadrata	Kvadrat očekivanja	F statistika	p-vrijednost
Godina	4	323416	80854	6.312	5.51e-05
Residuali	648	8301247	12811	-	-

Tablica 1.25: ANOVA test za bodove prijemnog ispita

Iz tablice 1.25 vidljivo je da na nivou značajnosti $\alpha = 0.05$ postoji statistički značajna razlika između barem dvije grupe. Provodimo Tukey test i dobivamo rezultate dane u tablici 1.26.

godina	razlika srednjih vrijednosti	95% interval pouzdanosti	p-vrijednost
2006-2005	−46.3674497	$\langle -83.680308, -9.054591 \rangle$	0.0064227
2007-2005	−19.6108696	$\langle -59.390981, 20.169241 \rangle$	0.6608409
2008-2005	−51.2926357	$\langle -89.918969, -12.666303 \rangle$	0.0027931
2009-2005	0.1287879	$\langle -38.278309, 38.535885 \rangle$	1.0000000
2007-2006	26.7565801	$\langle -11.673976, 65.187136 \rangle$	0.3158772
2008-2006	−4.9251860	$\langle -42.160170, 32.309798 \rangle$	0.9963358
2009-2006	46.4962375	$\langle 9.488731, 83.503744 \rangle$	0.0056378
2008-2007	−31.6817661	$\langle -71.388841, 8.025309 \rangle$	0.1875617
2009-2007	19.7396574	$\langle -19.754181, 59.233496 \rangle$	0.6489325
2009-2008	51.4214235	$\langle 13.089979, 89.752869 \rangle$	0.0024338

Tablica 1.26: Tukey test razlike uzoračkih očekivanja grupa

Iz tablice 1.26 vidimo da postoji statistički značajna razlika (na nivou značajnosti $\alpha = 0.05$) između 2006. i 2005., 2008. i 2005., 2009. i 2006. godine , te između 2009. i 2008. godine što se moglo predvidjeti iz grafičkog prikaza 1.28.

Testiramo još i jednakost uzoračkih očekivanja $\mu_{2010} = \mu_{2011}$, odnosno testiramo hipotezu

$$H_0 : \mu_{2010} = \mu_{2011}. \tag{1.25}$$

Testna statistika koju koristimo je

$$Z_{a,b} = \frac{\mu_a - \mu_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}, \tag{1.26}$$

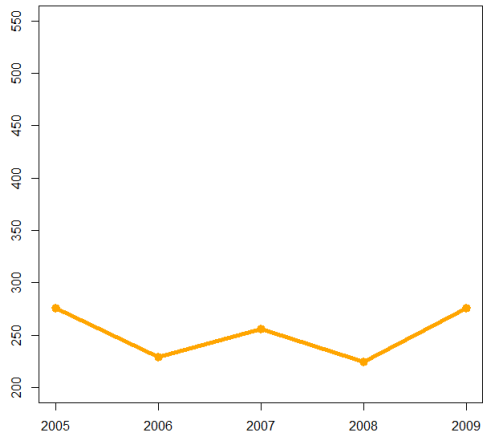
te pod pretpostavkom da vrijedi hipoteza H_0 imamo $Z \overset{H_0}{\sim} AN(0,1)$, odnosno uz nultu hipotezu o jednakostima očekivanja imamo da $Z \xrightarrow{\mathcal{D}} N(0,1)$ kada $n_a,\ n_b \longrightarrow +\infty$. Prema tome, za dovoljno velike n_g računamo pripadnu p-vrijednost kao $2 \cdot (\mathbb{P}(|Z| > z)) = 2 \cdot (1 - \mathbb{P}(|Z| \leq z))$.

Dobivena p- vrijednost za testnu hipotezu

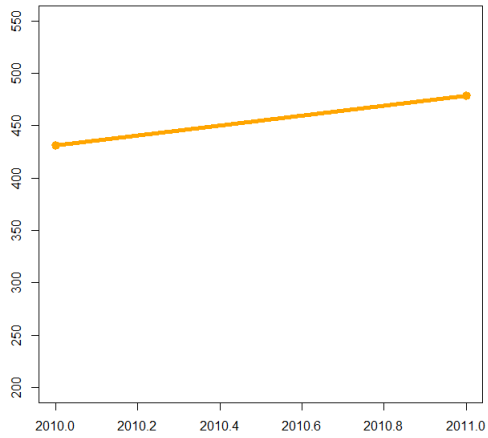
$$H_0 : \mu_{2010} = \mu_{2011} \tag{1.27}$$

je izrazito mala, dakle postoji statistički značajna razlika između μ_{2010} i μ_{2011} (na nivou značajnosti $\alpha = 0.05$).

Na slici 1.29 i 1.30 dan je grafički prikaz srednjih vrijednosti bodova iz škole za neuspješne studente.



Slika 1.29: Srednje vrijednosti bodova prijemnog za neuspješne studente



Slika 1.30: Srednje vrijednosti bodova prijemnog za neuspješne studente

Usporedimo li grafičke prikaze 1.19 i 1.20 s grafičkim prikazima 1.24, 1.25, 1.29 i 1.30 vidljivo je da uspješni studenti imaju u prosjeku više bodova ostvarenih na prijemnom ispitu (državnoj maturi), dok neuspješni studenti u prosjeku imaju manje bodova ostvarenih na prijemnom ispitu (državnoj maturi) od prosjeka ostvarenih bodova svih studenata.

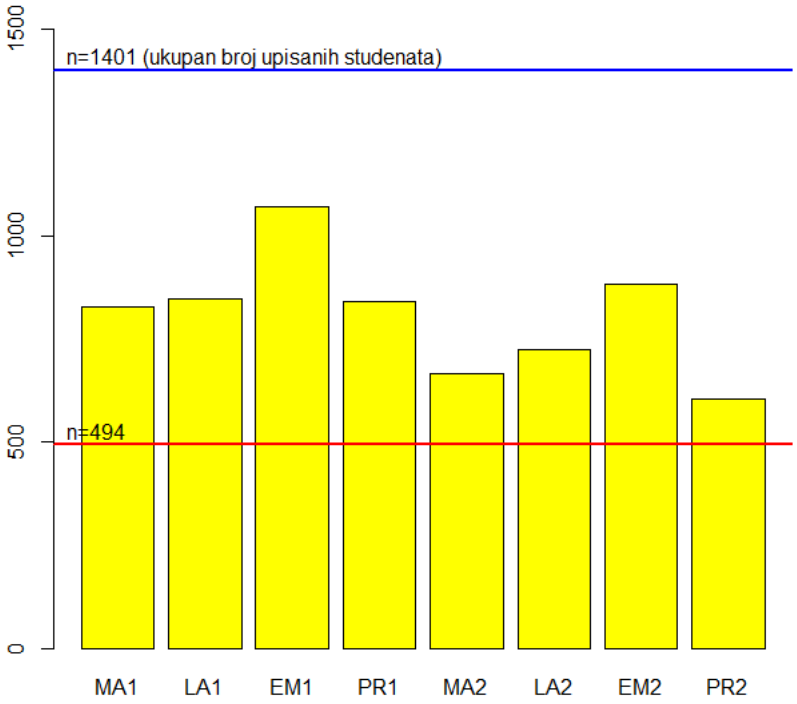
1.4 Analiza uspješnih studenata

Uspješnim studentom smatramo jedinku iz populacije koja je sakupila točna 60 ECTS bodova kroz prva dva semestra studiranja, odnosno jedinku koja je položila sve predmete prve godine u prvoj godini studiranja. U tablici 1.27 dan je prikaz položenosti pojedinih kolegija i ukupni udio uspješnih studenata u odnosu na promatrane godine za sve upisane studente.

Godina	ukupno	MA1	LA1	EM1	PR1	MA2	LA2	EM2	PR2	uspješni	%
2005	227	151	135	189	162	121	125	133	115	99	43.612
2006	223	142	134	162	147	110	82	116	82	74	33.184
2007	179	122	104	132	113	93	88	106	79	64	35.754
2008	181	83	85	106	86	74	79	90	64	52	28.729
2009	196	103	121	139	104	85	100	128	75	64	32.653
2010	198	105	135	165	120	86	114	149	102	78	39.394
2011	197	122	134	177	109	97	127	160	88	63	31.980
ukupno	1401	828	848	1070	841	666	725	882	605	494	35.261

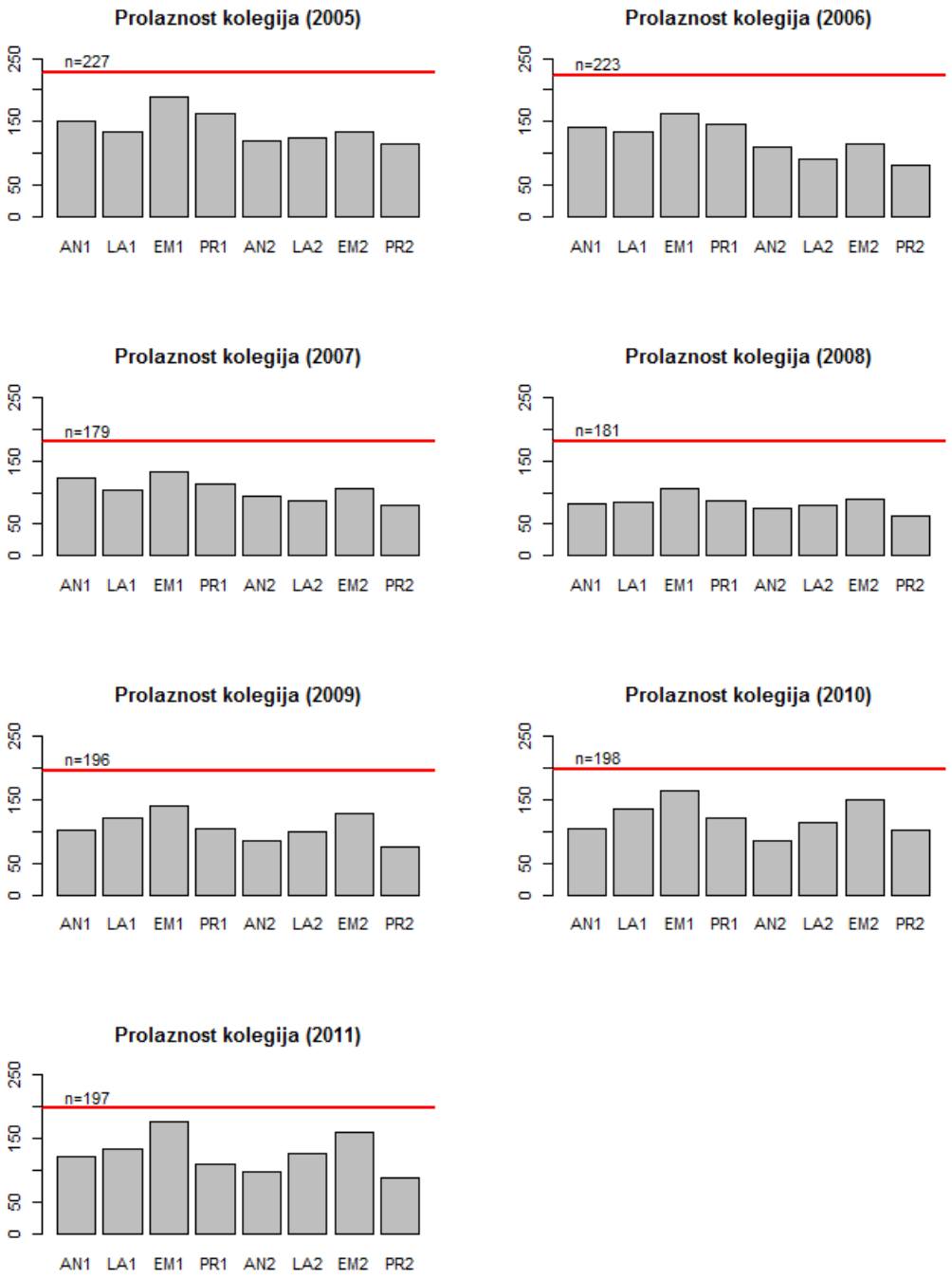
Tablica 1.27: Udio položenih kolegija i uspješnih studenata po godinama

Prvo dajemo grafički prikaz ukupne prolaznosti kolegija (sve godine zajedno) na slici 1.31 za sve studente. Plava linija označava ukupan broj upisanih studenata svih promatranih godina, dok crvena linija označava ukupan broj uspješnih studenata svih promatranih godina. S grafičkog prikaza vidljivo je da su kolegiji s najvećim udjelom položenosti Elementarna matematika 1 i Elementarna matematika 2, dok najmanje studenata polaže kolegije Programiranje 2 i Matematička analiza 2. Budući su kolegiji prvog semestra kolegiji prethodnici kolegijima drugog semestra, jasno je da će položenosti u drugom semestru biti manja nego u prvom. Nadalje, vidljivo je da se frekvencije položenosti prvog semestra (u obliku po kolegijima) ponašaju vrlo slično frekvencijama položenosti drugog semestra.



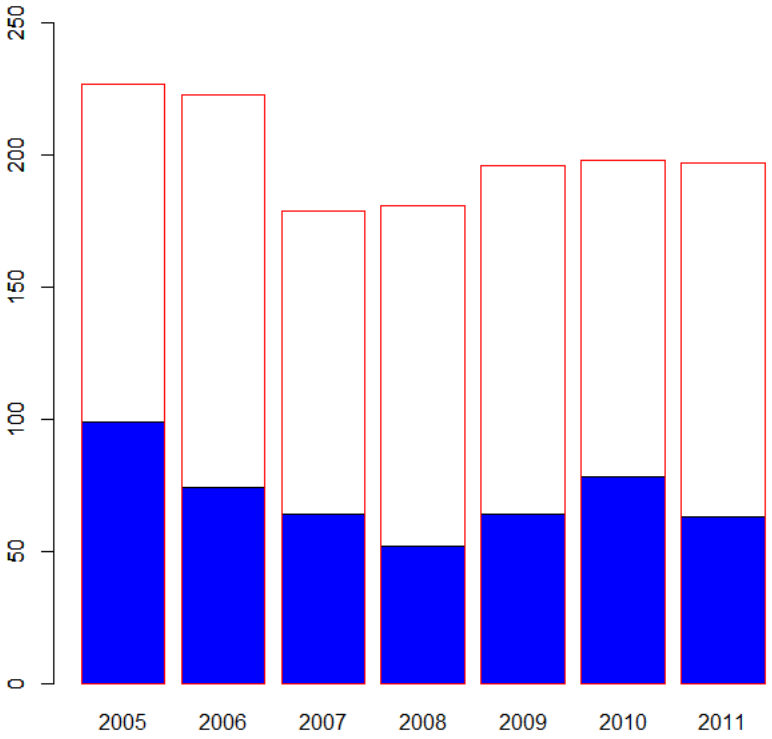
Slika 1.31: Ukupna prolaznosti kolegija

Na slici 1.32 dan je grafički prikaz prolaznosti kolegija po godinama za sve upisane studente, odnosno grafički prikaz ostatka tablice 1.27. Vidljivo je da je prolaznost kolegija Elementarna matematika 1 svake godine najveća, dok je prolaznost kolegija Programiranje 1 veća za generacije upisane 2005. i 2006. godine.



Slika 1.32: Prolaznost po kolegijima kroz godine

Također, dajemo i grafički prikaz odnosa upisanih i uspješnih studenata kroz godine na slici 1.33.



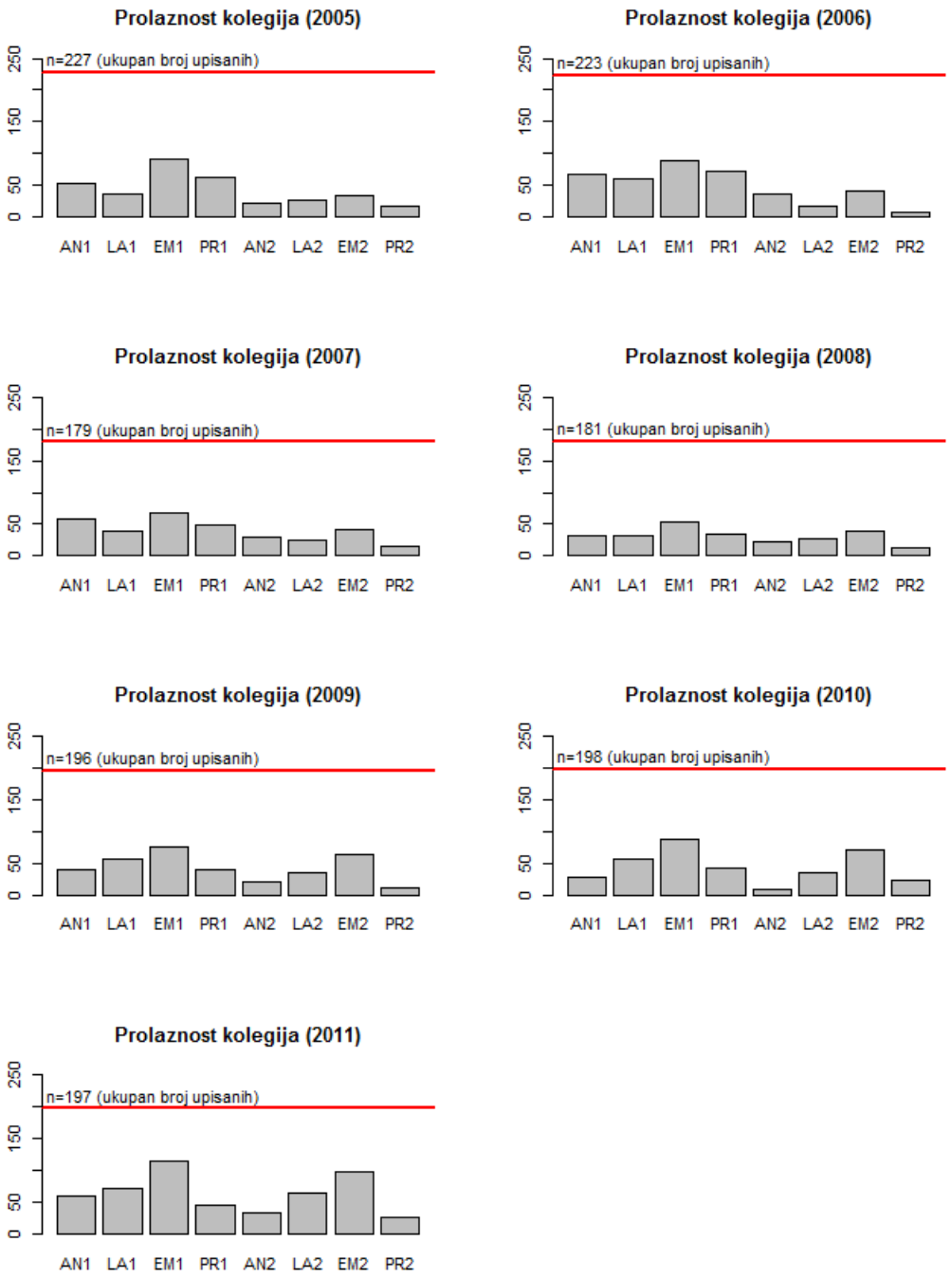
Slika 1.33: Odnos upisanih i uspješnih studenata kroz godine

Budući uspješni studenti imaju položene sve kolegije u prva dva semestra studiranja, promotrimo sada samo neuspješne studente i frekvencije položenih kolegija za neuspješne studente. Podaci su dani u tablici 1.28

Godina	neuspješni	MA1	LA1	EM1	PR1	MA2	LA2	EM2	PR2	ukupno	%
2005	128	52	36	90	63	22	26	34	16	227	56.388
2006	149	68	60	88	73	36	18	42	8	223	66.816
2007	115	58	40	68	49	29	24	42	15	179	64.246
2008	129	31	33	54	34	22	27	38	12	181	71.271
2009	132	39	57	75	40	21	36	64	11	196	67.347
2010	120	27	57	87	42	8	36	71	24	198	60.606
2011	134	59	71	114	46	34	64	97	25	197	68.020
ukupno	907	828	848	1070	841	666	725	882	605	494	64.739

Tablica 1.28: Udio položenih kolegija za neuspješne studenata po godinama

Na slici 1.34 dan je grafički prikaz prolaznosti kolegija po godinama za neuspješne studente, odnosno grafički prikaz tablice 1.28. Vidljivo je da je prolaznost kolegija Elementarna matematika 1 svake godine najveća, dok je prolaznost kolegija Programiranje 1 veća za generacije upisane 2005. i 2006. godine. Usporedimo li grafičke prikaze 1.32 i 1.34, vidljivo je da frekvencije imaju sličan oblik po predmetima, ali da je udio položenih kolegija kod neuspješnih studenata znatno manji nego kod cijele populacije.



Slika 1.34: Prolaznost po kolegijima za neuspješne studente

Neka je sada $X_g,\ g \in G = \{2005, 2006, 2007, 2008, 2009, 2010, 2011\}$ broj uspješnih studenata u godini g, n_g ukupni broj upisanih studenata u godini g i neka je p_g proporcija uspješnih studenata u godini g, $p_g = \frac{X_g}{n_g}$. Standardna pogreška proporcije p_g, u oznaci $\mathbf{SE}(p_g)$ je

$$\mathbf{SE}(p_g) = \sqrt{\frac{p_g(1-p_g)}{n_g}}, \tag{1.28}$$

dok je $100 \cdot (1-\alpha)\%$ interval pouzdanosti za proporciju p_g oblika

$$\left\langle p_g - z_{\frac{\alpha}{2}} \sqrt{\frac{p_g(1-p_g)}{n}}, p_g + z_{\frac{\alpha}{2}} \sqrt{\frac{p_g(1-p_g)}{n}} \right\rangle, \tag{1.29}$$

gdje je $z_{\frac{\alpha}{2}}$
Neka je X slučajna varijabla koja opisuje broj uspješnih studenata u godini. Tada je X suma Bernoullijevih slučajnih varijabli, odnosno X ima binomnu razdiobu, $X \sim B(p, n)$, gdje je $n = \sum_{g=2005}^{2011} n_g$ = ukupni broj upisanih studenata. Točkovni procjenitelj binomnog parametra p je

$$\hat{p} = \frac{X_{2005} + X_{2006} + X_{2007} + X_{2008} + X_{2009} + X_{2010} + X_{2011}}{n_{2005} + n_{2005} + n_{2006} + n_{2007} + n_{2008} + n_{2009} + n_{2010} + n_{2011}} = \frac{X}{n}. \tag{1.30}$$

Standardna pogreška procjenitelja $\hat{p}$, u oznaci $\mathbf{SE}(\hat{p})$ je

$$\mathbf{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \tag{1.31}$$

dok je $100 \cdot (1-\alpha)\%$ interval pouzdanosti za procjenitelj $\hat{p}$ oblika

$$\left\langle \hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right\rangle, \tag{1.32}$$

gdje je $z_{\frac{\alpha}{2}} \in \mathbb{R}^+$ takav da vrijedi $\mathbb{P}(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$, gdje je $Z \sim N(0,1)$.
Dobiveni rezultati nalaze se u tablici 1.29.

g	n_g	X_g	p_g	95% interval pouzdanosti za p_g	**SE**(p_g)
2005	227	99	0.43612	$\langle 0.3716126, 0.5006341\rangle$	0.032914
2006	223	74	0.33184	$\langle 0.2700370, 0.3936402\rangle$	0.031532
2007	179	64	0.35754	$\langle 0.2873305, 0.4277533\rangle$	0.035823
2008	181	52	0.28729	$\langle 0.2213714, 0.3532143\rangle$	0.033634
2009	196	64	0.32653	$\langle 0.2608797, 0.3921816\rangle$	0.033496
2010	198	78	0.39394	$\langle 0.3258800, 0.4619988\rangle$	0.034725
2011	197	63	0.31980	$\langle 0.2546684, 0.3849255\rangle$	0.033229
ukupno	n	X	$\hat{p}$	95% interval pouzdanosti za $\hat{p}$	**SE**$(\hat{p})$
Σ	1401	494	0.35261	$\langle 0.327587, 0.3776236\rangle$	0.012765

Tablica 1.29: Proporcije za $X_g,\ g \in G$ i X

Želimo testirati jednakost pojedine proporcije p_g i binomnog parametra p. Koristimo test o proporciji za velike n ($n > 30$), odnosno testiramo

$$H_0 : p = p_g, \tag{1.33}$$

za svaki $g \in G$. Koristimo testnu statistiku

$$Z_g = \frac{\hat{p} - p_g}{\sqrt{\frac{p_g(1-p_g)}{n}}} = \frac{X - np_g}{\sqrt{np_g(1-p_g)}}, \tag{1.34}$$

koja uz hipotezu H_0 ima standardnu normalnu razdiobu, $Z_g \overset{H_0}{\sim} N(0,1)$. Pripadnu p-vrijednost računamo kao

$$p_g = 2 \cdot \mathbb{P}(Z > |z_g|),\ g \in G\ (Z \sim N(0,1)), \tag{1.35}$$

te dobivamo rezultate testa o proporciji nalaze se u tablici 1.30.

g	n_g	p_g	Z_g	p-vrijednost
2005	227	0.43612	-6.303805	0.0000000003
2006	223	0.33184	1.650756	0.0987884894
2007	179	0.35754	-0.385534	0.6998418307
2008	181	0.28729	5.402530	0.0000000657
2009	196	0.32653	2.081215	0.0374142072
2010	198	0.39394	-3.166321	0.0015438031
2011	197	0.31980	2.632975	0.0084640611

Tablica 1.30: Rezultati testa o proporciji

Iz tablice 1.30 zaključujemo da na nivou značajnosti $\alpha = 0.05$ odbacujemo hipotezu o jednakosti H_0 u svim slučajevima osim za 2006. i 2007. godinu.

Neka su $a, b \in G$. Testiramo hipotezu o jednakosti parametara za godine a i b. odnosno

$$H_0 : p_a = p_b. \tag{1.36}$$

Definiramo

$$p_{a,b} := \frac{X_a + X_b}{n_a + n_b} \ (= p_{b,a}), \tag{1.37}$$

te računamo testnu statistiku

$$Z_{a,b} = \frac{\hat{p}_a - \hat{p}_b}{\sqrt{p_{a,b}(1 - p_{a,b})}\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}} \overset{H_0}{\sim} AN(0, 1) \tag{1.38}$$

g	2005	2006	2007	2008	2009	2010	2011
2005	-	0.02299	0.10878	0.00198	0.02092	0.37887	0.01395
2006	0.02299	-	0.58959	0.33649	0.90817	0.18547	0.79279
2007	0.10878	0.58959	-	0.15385	0.52699	0.46642	0.43959
2008	0.00198	0.33649	0.15385	-	0.40954	0.02892	0.49261
2009	0.02092	0.90817	0.52699	0.40954	-	0.16349	0.88652
2010	0.37887	0.18547	0.46642	0.02892	0.16349	-	0.12409
2011	0.01395	0.79279	0.43959	0.49261	0.88652	0.12409	-

Tablica 1.31: Rezultati testa o jednakosti poroporcija po godinama

Iz tablice 1.31 vidljivo je da na nivou značajnosti $\alpha = 0.05$ postoji statistički značajna razlika između p_{2005} i p_{2006}, p_{2005} i p_{2008}, p_{2005} i p_{2009}, p_{2005} i p_{2011}, te između p_{2008} i p_{2010}. Na nivou značajnosti $\alpha = 0.01$ postoji statistički značajna razlika jedino između p_{2005} i p_{2008}.

Poglavlje 2

Logistička regresija: univarijatni model

2.1 Uvod u logističku regresiju

U modernoj statistici metode regresije postale su sastavni dio svake statističke analize podataka koja opisuje odnos između zavisne varijable i jedne ili više nezavisnih varijabli. Zavisnu varijablu zovemo varijabla odaziva i označavamo sa Y, dok nezavisne varijable zovemo varijable poticaja (ponekad i kovarijate) i označavamo sa $X_1, X_2, \ldots, X_k$, gdje je $k \in \mathbb{N}$ broj varijabli poticaja.

Najpoznatije metode regresije su linearna regresija, logistička regresija i Poissonova regresija.

Definicija 2.1.1. *Uvjetno očekivanje varijable odaziva Y uz danu vrijednost varijable poticaja $X = x$ označavamo s $\mathbb{E}(Y|X = x)$.*

U linearnoj regresiji za dani skup varijabli poticaja tražimo najbolju hiperravninu koja opisuje varijablu odaziva (u slučaju $n = 1$ dobivamo pravac). Pretpostavljamo da uvjetno očekivanje $\mathbb{E}(Y|X = x)$ možemo izraziti linearno preko x (ili preko neke transformacije od x), odnosno da bez smanjenja općenitosti vrijedi

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x, \tag{2.1}$$

gdje su $\beta_0, \beta_1 \in \mathbb{R}$ parametri. Također pretpostavljamo da je varijabla odaziva u linearnoj regresiji neprekidna.

Vrlo često su varijable odaziva diskretne, te poprimaju prebrojivo ili čak konačno vrijednosti. Modeli logističke regresije bave se upravo takvom vrstom statističke analize. Kao i u svakom modelu regresije, cilj je pronaći što bolji i razumniji model pomoću kojega opisujemo odnos između varijable odaziva i jedne ili više varijabli poticaja.

Definicija 2.1.2. *Varijablu odaziva Y zovemo dihotomna ili binarna ukoliko poprima samo dva moguća stanja. Takve varijable u pravilu kodiramo s dva stanja: $y = 0$ ili $y = 1$.*

Napomena 2.1.3. *Ukoliko je Y dihotomna varijabla, raspisivanjem uvjetnog očekivanja dobivamo*

$$\mathbb{E}(Y|X = x) = 0 \cdot \mathbb{P}(Y = 0|X = x) + 1 \cdot \mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 1|X = x), \tag{2.2}$$

što nam daje interpretaciju uvjetnog očekivanja zavisne varijable Y uz dano $X = x$ kao uvjetne vjerojatnosti da zavisna varijabla Y poprimi vrijednost 1 uz dano $X = x$.

Iz napomene 2.1.3 slijedi da dihotomna varijabla odaziva $Y|X = x$ ima Bernoullijevu radiobu, odnosno da vrijedi

$$Y|X = x \sim \begin{pmatrix} 0 & 1 \\ 1 - \mathbb{E}(Y|X = x) & \mathbb{E}(Y|X = x) \end{pmatrix}. \tag{2.3}$$

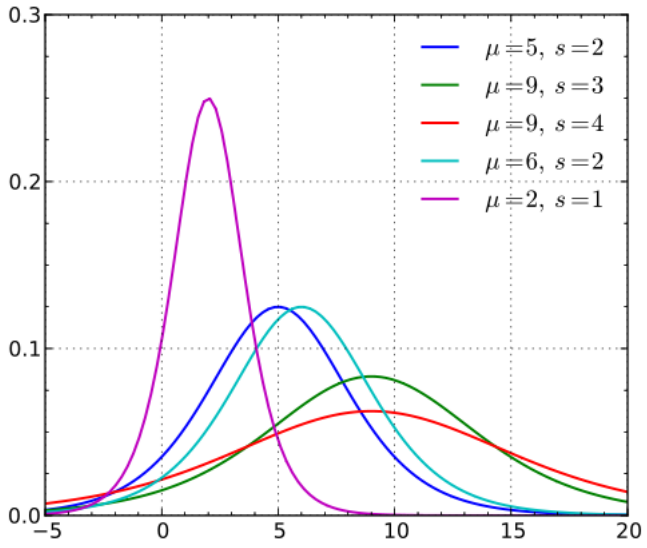
U slučaju kada je Y dihotomna varijabla $\mathbb{E}(Y|X = x)$ nalazi se u segmentu $[0, 1]$, a vidjet ćemo da se $\mathbb{E}(Y|X = x)$ postepeno približava 0 (s desna), odnosno 1 (s lijeva). Krivulje takvog oblika nazivamo S-oblikovane krivulje (podsjećaju na kumulativne funkcije distribucije neprekidnih slučajnih varijabli). Model koji ćemo koristiti za opis $\mathbb{E}(Y|X = x)$ u slučaju dihotomne varijable odaziva Y bazira se na logističkoj distribuciji.

Logistička distribucija

U teoriji vjerojatnosti i statistici, logistička distribucija je neprekidna vjerojatnosna distribucija. Vjerojatnostna funkcija gustoće logističke distribucije $f : \mathbb{R} \to [0, 1]$ dana je formulom

$$f(x; \mu, s) = \frac{e^{-\frac{x-\mu}{s}}}{s(1 + e^{-\frac{x-\mu}{s}})^2}, \ x \in \mathbb{R}, \tag{2.4}$$

gdje su $\mu, s \in \mathbb{R}$, $s > 0$ zadani parametri. Na slici 2.1 prikazani su oblici funkcije gustoće logističke distribucije u ovisnosti o parametrima μ i s. Vjerojatnosna funkcija gustoće



Slika 2.1: Oblici funkcije gustoće logističke distribucije

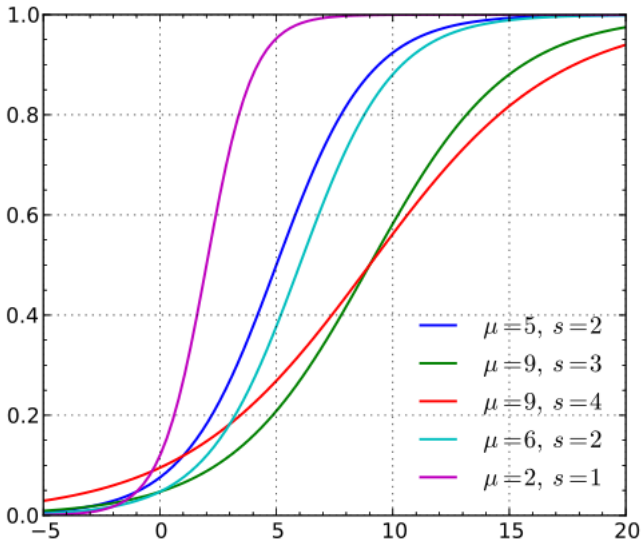
izgledom podsjeća na normalnu distribuciju, ali ima teže repove (veći koeficijent zaobljenosti; kurtosis).

Kumulativna funkcija distribucije logističke distribucije $F : \mathbb{R} \to [0, 1]$ naziva se logistička funkcija. Formula logističke funkcije je oblika

$$F(x; \mu, s) = \int_{-\infty}^{x} f(x; \mu, s)dx = \frac{1}{1 + e^{-\frac{x-\mu}{s}}}, \ x \in \mathbb{R}, \tag{2.5}$$

za zadane parametre $\mu, s \in \mathbb{R}, \ s > 0$.

Logistička funkcija se koristi u logističkoj regresiji i umjetnim neuronskim mrežama (artificial neural networks; ANNs). Na slici 2.2 prikazani su oblici logističke funkcije u ovisnosti o parametrima μ i s.



Slika 2.2: Oblici logističke funkcije

U tablici 2.1 se nalaze osnovna svojstva logističke distribucije.

očekivanje	medijan	varijanca	standardna devijacija	skewness	kurtosis
μ	μ	$\frac{s^2\pi^2}{3}$	$\frac{s\pi}{\sqrt{3}}$	0	1.2

Tablica 2.1: Osnovna svojstva logističke distribucije

Opće oznake

Napomena 2.1.4. *Kako bi pojednostavili notaciju, uvodimo oznaku* $\pi(x) = \mathbb{E}(Y|X = x)$ *za uvjetno očekivanje od* Y *uz dano* $X = x$ *kada koristimo logističku distribuciju.*

Ukoliko u formuli logističke funkcije (formula 2.5 stavimo supstituciju $\mu = \frac{\beta_1}{\beta_0}$ i $s = \frac{1}{\beta_0}$ dobivamo općenitu formulu za uvjetno očekivanje od Y uz dano $X = x$, odnosno vrijedi

$$\pi(x) = F(x; \mu, s) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{2.6}$$

Nadalje, uvodimo transformaciju uvjetnog očekivanja $\pi(x)$ u oznaci $g(x) := G(\pi(x))$, gdje je $G : \mathbb{R} \to \mathbb{R}$ definirana kao

$$G(x) = \ln\left(\frac{x}{1-x}\right). \tag{2.7}$$

Preslikavanje g nazivamo logaritamska transformacija (logit transformation) uvjetnog očekivanja $\pi(x)$ i dobivamo

$$g(x) = G(\pi(x)) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \ln\left(\frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{1 - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}\right) = \beta_0 + \beta_1 x \tag{2.8}$$

Napomena 2.1.5. *Preslikavanje* g *ima sva bitna svojstva potrebna za linearni regresijski model: linearno je, može biti neprekidno i slika preslikavanja* g *je podskup realnih brojeva,* $\mathrm{Im}(g) \subset \mathbb{R}$ *u ovisnosti o domeni nezavisne varijable.*

Neka je $\varepsilon = y - \pi(x)$ odstupanje opservacije od uvjetnog očekivanja uz dano $X = x$ ($\varepsilon := \varepsilon|X = x$). Budući je Y dihotomna varijabla (ona poprima vrijednosti 0 ili 1) i zbog napomene 2.1.3 dobivamo dva slučaja:

i) $Y = 0 \implies \varepsilon|X = x = -\pi(x)$ s vjerojatnošću $\mathbb{P}(Y = 0|X = x) = 1 - \pi(x)$

ii) $Y = 1 \implies \varepsilon|X = x = 1 - \pi(x)$ s vjerojatnošću $\mathbb{P}(Y = 1|X = x) = \pi(x)$.

Budući da za uvjetno očekivanje i uvjetnu varijancu greške ϵ uz dano $X = x$ vrijedi

$$\mathbb{E}(\varepsilon|X = x) = -\pi(x)(1 - \pi(x)) + (1 - \pi(x))\pi(x) = 0 \tag{2.9}$$

i

$$Var(\varepsilon|X = x) = \pi(x)(1 - \pi(x)) \tag{2.10}$$

respektivno, slijedi da greška ϵ uz dano $X = x$ ima Bernoullijevu radiobu, odnosno da vrijedi

$$\varepsilon|X = x \sim \begin{pmatrix} -\pi(x) & 1 - \pi(x) \\ 1 - \pi(x) & \pi(x) \end{pmatrix}. \tag{2.11}$$

Napomena 2.1.6. *Kako iz relacije 2.10 slijedi nehomogenost varijance greške ε neće vrijediti Gauss-Markovljevi uvjeti.*

2.2 Prilagodba modela logističke regresije

Neka je X nezavisna varijabla, Y zavisna dihotomna varijabla i neka je dani uzorak on $n \in \mathbb{N}$ nezavisnih opservacija (x_i, y_i), $i \in \{1, 2, \ldots, n\}$. Iz relacije 2.8 slijedi kako je logaritamska transformacija uvjetnog očekivanja $\pi(x)$ oblika

$$g(x) = \beta_0 + \beta_1 x. \tag{2.12}$$

Želimo na temelju danog uzorka (x_i, y_i), $i \in \{1, 2, \ldots, n\}$ procijeniti vektor parametara $\boldsymbol{\beta} := (\beta_0, \beta_1)$.

U linearnim regresijskim modelima najbolji, linearni, nepristrani procjenitelj (BLUE) za $\boldsymbol{\beta} := (\beta_0, \beta_1)$ je LS-procjenitelj $\hat{\boldsymbol{\beta}} = (\hat{\beta_0}, \hat{\beta_1})$, odnosno procjenitelj dobiven metodom najmanjih kvadrata (least squares method).

Kako iz napomene 2.1.6 slijedi da u modelu s dihotomnom varijablom odaziva Y neće vrijediti Gauss-Markovljevi uvjeti , ne možemo primijeniti metodu najmanjih kvadrata.

Općenita metoda procjene za vektor parametara $\boldsymbol{\beta}$ je metoda maksimalne vjerodostojnosti.

Definicija 2.2.1. *Pod pretpostavkom da su opažanja nezavisna definiramo funkciju vjerodostojnosti, u oznaci ℓ, kao*

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n} \xi(x_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x))^{1-y_i}. \tag{2.13}$$

Nadalje, za funkciju vjerodostojnosti ℓ definiramo funkciju log-vjerodostojnosti, u oznaci $\mathcal{L}$, kao

$$\mathcal{L}(\boldsymbol{\beta}) = \ln\left(\ell(\boldsymbol{\beta})\right) = \sum_{i=1}^{n} \Big(y_i \ln\left(\pi(x_i)\right) + (1 - y_i) \ln\left(1 - \pi(x_i)\right)\Big) \tag{2.14}$$

Napomena 2.2.2. *Princip maksimalne vjerodostojnosti procjenjuje vektor parametara $\boldsymbol{\beta}$ s vektorom parametara $\hat{\boldsymbol{\beta}}$ koji maksimizira funkciju vjerodostojnosti, ondosno s $\hat{\boldsymbol{\beta}}$ za koji vrijedi*

$$\ell(\hat{\boldsymbol{\beta}}) = \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}). \tag{2.15}$$

Kako je matematički jednostavnije maksimizirati funkciju log-vjerodostojnosti tražimo $\hat{\boldsymbol{\beta}}$ za koji vrijedi

$$\mathcal{L}(\hat{\boldsymbol{\beta}}) = \max_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}). \tag{2.16}$$

Parcijalno defiriviramo funkciju log-vjerodostojnosti $\mathcal{L}$ u odnosu na β_0 i β_1 i izjednačavamo s nulom. Jednadžbe koje dobivamo zovemo jedndžbe vjerodostojnosti i one su

$$\frac{\partial}{\partial \beta_0} \mathcal{L}(\beta_0, \beta_1) = 0 \iff \sum_{i=1}^{n} (y_i - \pi(x_i)) = 0 \tag{2.17}$$

i

$$\frac{\partial}{\partial \beta_1}\mathcal{L}(\beta_0, \beta_1) = 0 \iff \sum_{i=1}^{n} x_i(y_i - \pi(x_i)) = 0 \tag{2.18}$$

Jednadžbe vjerodostojnosti 2.17 i 2.18 su nelinearne u parametrima β_0 i β_1. Metode za rješavanje tih jednadžbi su iterativne i provode se pomoću računala (iterativna težinska metoda najmanjih kvadrata). Detaljniji pogled na rješavanje jednadžbi vjerodostojnosti dan je u dodatku rada.

Napomena 2.2.3. *Iz relacije 2.17 dobivamo da je suma opserviranih vrijednosti y_i jednaka sumi procijenjenih očekivanih vrijednosti $\hat{\pi}(x_i)$, odnosno vrijedi*

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i), \tag{2.19}$$

gdje je $\hat{\pi}(x_i)$ procjena maksimalne vjerodostojnosti uvjetnog očekivanja $\pi(x_i)$, odnosno

$$\hat{\pi}(x_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}. \tag{2.20}$$

2.3 Testiranje značajnosti koeficijenata

U prethodnom poglavlju pokazali smo kako procijeniti logistički model za dani uzorak opažanja (x_i, y_i), $i = 1, \dots n$. Želimo testirati statističku hipotezu o značajnosti utjecaja (u nekom smislu) nezavisne varijable na zavisnu. Ukoliko su procijenjene vrijednosti u modelu s nezavisnom varijablom točnije (u nekom smislu) nego u modelu bez te varijable, onda kažemo da je nezavisna varijabla značajna.

Test omjera vjerodostojnosti

U modelu logističke regresije usporedba dvaju modela provodi se pomoću funkcije vjerodostojnosti koju smo definirali u definiciji 2.2.1, a testna statistika je oblika

$$D = -2\ln\left(\frac{vjerodostojnost\ procijenjenog\ modela}{vjerodostojnost\ saturiranog\ modela}\right) \overset{oznaka}{=} -2\ln\left(\frac{\ell(\mathcal{P})}{\ell(\mathcal{S})}\right), \tag{2.21}$$

gdje je saturirani ili zasićeni model onaj koji sadrži jednak broj nepoznatih parametara kolika je veličina uzorka, odnosno u kojem vrijedi

$$\hat{\pi}(x_i) = y_i. \tag{2.22}$$

Izraz unutar velikih zagrada ($e^{\frac{-D}{2}}$) zovemo omjer vjerodostojnosti. Test koji provodimo zovemo test omjera vjerodostojnosti, a testnu statistiku D odstupanje ili devijanca.

Propozicija 2.3.1. $\ell(\mathcal{S}) = 1$

Dokaz. Iz relacije 2.22 slijedi da je

$$\ell(\mathcal{S}) = \prod_{i=1}^{n} \hat{\pi}(x_i)^{y_i}(1 - \hat{\pi}(x))^{1-y_i} = \prod_{i=1}^{n} y_i^{y_i}(1 - y_i)^{1-y_i}. \tag{2.23}$$

Kako je Y dihotomna varijabla, imamo da je $y_i \in \{0, 1\}$. Dobivamo dva slučaja:

i) $y_i = 0 \implies \xi(x_i) = \lim\limits_{y_i \to 0+} y_i^{y_i} \cdot 1 = \lim\limits_{a \to 0+} a^a$

ii) $y_i = 1 \implies \xi(x_i) = \lim\limits_{y_i \to 1-} 1 \cdot (1 - y_i)^{1-y_i} = \lim\limits_{a \to 0+} a^a$

Budući vrijedi

$$\lim_{a \to 0+} a^a = \lim_{a \to 0+} e^{\ln(a^a)} = \lim_{a \to 0+} e^{a\ln(a)} = e^{\lim\limits_{a \to 0+} a\ln(a)} = (L'Hospital) = e^0 = 1, \tag{2.24}$$

dobivamo

$$\ell(\mathcal{S}) \overset{BSO}{=} \prod_{i=1}^{n} \lim_{a \to 0+} a^a = 1 \tag{2.25}$$

□

Napomena 2.3.2. *Iz propozicije 2.3.1 dobivamo*

$$D = -2\ln\left(\frac{\ell(\mathcal{P})}{\ell(\mathcal{S})}\right) = -2\ln\left(\ell(\mathcal{P})\right) = -2\ln\left(\prod_{i=1}^{n}\hat{\pi}(x_i)^{y_i}(1-\hat{\pi}(x_i))^{1-y_i}\right), \tag{2.26}$$

odnosno

$$D = -2\sum_{i=1}^{n}\left(y_i\ln\left(\hat{\pi}(x_i)\right) + (1-y_i)\ln\left(1-\hat{\pi}(x_i)\right)\right) \tag{2.27}$$

Definiramo statistiku G kao razliku devijanci D u ovisnosti o tome sadrži li model zavisnu varijablu koju promatramo ili ne,

$$G = D(\textit{model bez varijable}) - D(\textit{model s varijablom}) \tag{2.28}$$

Napomena 2.3.3. *Zbog propozicije 2.3.1 statistiku G možemo izraziti kao*

$$G = -2\ln\left(\frac{\textit{vjerodostojnost modela bez varijabla}}{\textit{vjerodostojnost modela s varijablom}}\right) \tag{2.29}$$

Neka su n_0 i n_1 brojevi opažanja zavisne varijable s vrijednostima 0 i 1 respektivno, odnosno

$$n_0 = \sum_{i=1}^{n}(1-y_i),\ n_1 = \sum_{i=1}^{n}y_i.$$

U univarijatnom modelu bez varijable procjena maksimalne vjerodostojnosti za parametar β_0 je $\hat{\beta}_0 = \ln(\frac{n_1}{n_0})$. U tom slučaju, uz oznaku $\hat{\pi}(x_i) = \hat{\pi}_i$ dobivamo

$$G = -2\ln\left(\frac{\left(\frac{n_0}{n}\right)^{n_0}\cdot\left(\frac{n_1}{n}\right)^{n_1}}{\prod_{i=1}^{n}\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{1-y_i}}\right), \tag{2.30}$$

odnosno

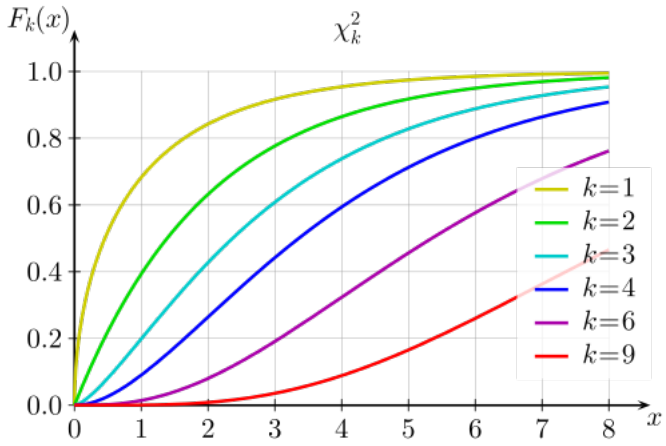
$$G = 2\cdot\left(\sum_{i=1}^{n}\left(y_i\ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)\right) - \left(n_1\ln(n_1) + n_0\ln(n_0) - n\ln(n)\right)\right) \tag{2.31}$$

Napomena 2.3.4. *Pod pretpostavkom da su n i* $\min(n_0, n_1)$ *dovoljno veliki i pod hipotezom da je* $\beta_1 = 0$*, statistika G ima* χ^2 *razdiobu s jednim stupnjem slobode, odnosno vrijedi*

$$H_0 : \beta_1 = 0 \implies G \overset{H_0}{\sim} \chi^2(1) \tag{2.32}$$

Na slici 2.3 nalaze se vjerojatnosne funkcije distribucije χ^2 razdiobe u ovisnosti o broju stupnjeva slobode, a kao primjer računanja p vrijednosti uzimamo $G = 29.31$ i dobivamo

$$p = \mathbb{P}(\chi^2(1) > 29.31) = 1 - \mathbb{P}(\chi^2(1) \leq 29.31) = 1 - F_{\chi^2(1)}(29.31) < 0.001$$



Slika 2.3: Oblici funkcije distribucije $\chi^2(k)$ razdiobe

Waldov test

Pretpostavljamo da su n i $\min(n_0, n_1)$ dovoljno veliki kao u napomeni 2.3.4. Waldovu testnu statistiku dobivamo kao omjer procjene maksimalne vjerodostojnosti parametra nagiba, $\hat{\beta}_1$ i procjene njegove standardne pogreške. Označavamo ju s W i dobivamo

$$W = \frac{\hat{\beta}_1}{\widehat{\mathbf{SE}}(\hat{\beta}_1)} \overset{H_0}{\sim} N(0,1), \tag{2.33}$$

gdje je $H_0 : \beta_1 = 0$

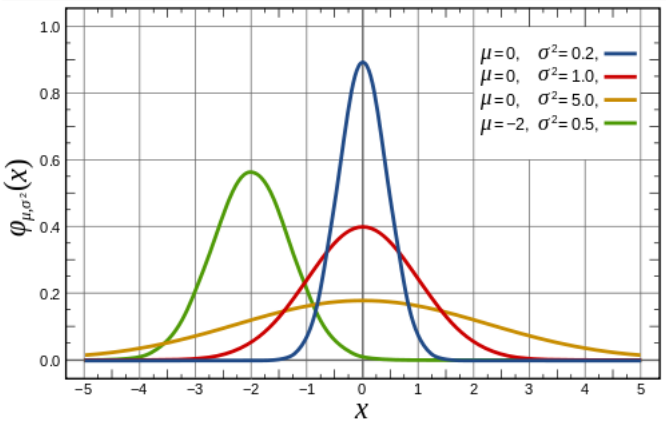
Na slici 2.4 nalaze se oblici vjerojatnosne funkcije gustoće $N(\mu, \sigma^2)$ razdiobe u ovisnosti o parametrima μ i σ^2, a kao primjer računanja p vrijednosti uzimamo $W = 4.61$ i dobivamo

$$p = \mathbb{P}(|W| > 4.61) = 1 - \mathbb{P}(|W| \leq 4.61) = 1 - \int_{-4.61}^{4.61} f_{N(0,1)}(x)dx < 0.001 \tag{2.34}$$

Napomena 2.3.5. *Ponekad se umjesto* $W \overset{H_0}{\sim} N(0,1)$ *koristi činjenica* $W^2 \overset{H_0}{\sim} \chi^2(1)$. *Važno je napomenuti da Waldov test ponekad ne odbacuje nultu hipotezu iako je nezavisna varijabla značajna.*

Score test

Score test je test temeljen na teoriji distribucije derivacija funkcije log vjerodostojnosti i ne zahtjeva računanje procjene parametara. Općenito ga koristimo u multivarijatnom modelu.



Slika 2.4: Oblici funkcije gustoće $N(\mu, \sigma^2)$ razdiobe

Testna statistika (**ST**) je oblika

$$\mathbf{ST} = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sqrt{\overline{y}(1-\overline{y})\sum_{i=1}^{n}(x_i - \overline{x})^2}} \overset{H_0}{\sim} N(0,1) \tag{2.35}$$

Napomena 2.3.6. *U primjenama su testne statistike za test omjera vjerodostojnosti, Waldov test i Score test približno jednake. U slučaju kada dobivamo p vrijednost na granici zadane značajnosti koristimo sve testne statistike.*

2.4 Procjene intervala pouzdanosti

Za procijenjene parametre $\hat{\beta}_0$ i $\hat{\beta}_1$ i za zadanu razinu značajnosti α računamo $(1-\alpha)\cdot 100\%$ intervale pouzdanosti. Dobivamo ih na osnovi pripadnog Waldovog testa i zovemo ih intervali pouzdanosti na osnovi Walda. Rubne točke $(1-\alpha)\cdot 100\%$ intervala pouzdanosti su

$$\hat{\beta}_1 \pm z_{1-\alpha/2} \cdot \widehat{\mathbf{SE}}(\beta_1) \tag{2.36}$$

za parametar nagiba $\hat{\beta}_1$ i

$$\hat{\beta}_0 \pm z_{1-\alpha/2} \cdot \widehat{\mathbf{SE}}(\beta_0) \tag{2.37}$$

za parametar konstante $\hat{\beta}_0$, gdje je $z_{1-\alpha/2}$ gornja $(1-\alpha/2)\cdot 100\%$ točka standardne normalne distribucije, a $\widehat{\mathbf{SE}}(\cdot)$ procjena standardne pogreške za danu procjenu parametra.

Napomena 2.4.1. *Određivanje i detaljniji uvid u procjenu standardne pogreške* $\widehat{\mathbf{SE}}(\cdot)$ *za dani parametar dan je unutar poglavlja multivarijatne analize.*

Definiramo procjenu $\hat{g}$ za logaritamsku transformaciju uvjetnog očekivanja $\pi(x)$ kao

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{2.38}$$

te rubne točke $(1-\alpha)100\%$ intervala pouzdanosti na osnovi Walda kao

$$\hat{g}(x) \pm z_{1-\alpha/2} \cdot \widehat{\mathbf{SE}}(\hat{g}(x)), \tag{2.39}$$

gdje je $\widehat{\mathbf{SE}}(\hat{g}(x))$ pozitivni drugi korijen procjene varijance $\widehat{\mathbf{Var}}(\hat{g}(x))$.

Napomena 2.4.2. *Procjenitelj varijance procjene logaritamske transformacije računamo kao varijancu sume, odnosno*

$$\widehat{\mathbf{Var}}(\hat{g}(x)) = \widehat{\mathbf{Var}}(\hat{\beta}_0 + \hat{\beta}_1 x) = \widehat{\mathbf{Var}}(\hat{\beta}_0) + x^2\widehat{\mathbf{Var}}(\hat{\beta}_1) + 2\cdot\widehat{\mathbf{Cov}}(\beta_0, \beta_1). \tag{2.40}$$

Dobivamo procjenu uvjetnog očekivanje $\hat{\pi}(x)$ kao

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1+e^{\hat{g}(x)}}, \tag{2.41}$$

te rubne točke $(1-\alpha)100\%$ pouzdanih intervala na osnovi Walda za procjenu uvjetnog očekivanja kao

$$\frac{e^{\hat{g}(x)\pm z_{1-\alpha/2}\widehat{\mathbf{SE}}(\hat{g}(x))}}{1+e^{\hat{g}(x)\pm z_{1-\alpha/2}\widehat{\mathbf{SE}}(\hat{g}(x))}}. \tag{2.42}$$

Poglavlje 3

Logistička regresija: multivarijatni model

3.1 Uvod u multivarijatni model i prilagodba modela

U prethodnom poglavlju uveli smo pojam logističke regresije u slučaju kada imamo samo jednu nezavisnu varijablu poticaja. Neka je sada $p \in \mathbb{N}$ proizvoljan broj nezavisnih varijabli poticaja $X_1, \dots X_p$, te neka je Y zavisna dihotomna varijabla odaziva. Promatramo uvjetnu vjerojatnost ishoda $Y = 1$ uz dano $\mathbf{x}^T = (x_1, \dots, x_p)$, u oznaci $\pi(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{x})$. Logit transformacija multivarijatnog modela logističke regresije je oblika

$$g(\mathbf{x}) = \ln\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right) = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_p \cdot x_p, \tag{3.1}$$

gdje je za model multivarijatne logističke regresije dobivamo

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}. \tag{3.2}$$

Neka je $n \in \mathbb{N}$ i neka je $(\mathbf{x}_i, y_i),\ i = 1, \dots, n$ dani uzorak od n nezavisnih opažanja. Slično kao i u slučaju univarijatnog modela procijenjujemo vektor od $p + 1$ parametara oblika

$$\beta^T = (\beta_0, \beta_1, \dots, \beta_p). \tag{3.3}$$

Procjenu radimo metodom maksimalne vjerodostojnosti, te dobivamo slične jednadžbe vjerodostojnosti kao i u univarijatnom modelu koje su oblika

$$\sum_{i=1}^{n} (y_i - \pi(\mathbf{x}_i)) = 0 \tag{3.4}$$

i

$$\sum_{i=1}^{n} x_{ij}(y_i - \pi(\mathbf{x}_i)) = 0,\ j = 1, 2, \dots, p. \tag{3.5}$$

Uzimamo parcijalne derivacije drugog reda funkcije log-vjerodostojnosti i dobivamo

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \cdot \pi(\mathbf{x}_i) \cdot (1 - \mathbf{x}_i)) \tag{3.6}$$

i

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \cdot \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} \cdot x_{il} \cdot \pi(\mathbf{x}_i) \cdot (1 - \pi(\mathbf{x}_i)), \tag{3.7}$$

gdje su $j, l = 0, 1, 2, \dots, p$.

Definiramo $(p + 1) \cdot (p + 1)$ matricu informacije (observed informatin matrix), u oznaci $\mathbb{I}(\beta)$ kao matricu dobivenu negativnim vrijednostima gore navedenih parcijalnih derivacija

drugog reda. Kovarijaciona matrica procijenjenih parametara dobivena je kao inverz matice informacije, odnosno vrijedi

$$\mathbf{Var}(\beta) = \mathbb{I}^{-1}(\beta). \tag{3.8}$$

Eksplicitan oblik matrice $\mathbf{Var}(\beta)$ nemoguće je dobiti osim u vrlo specijalnim slučajevima, stoga se koristimo sljedećom notacijom: dijagonalni element kovarijacione matrice na mjestu j označavamo s $\mathbf{Var}(\beta_j)$ što je varijanca procjenitelja maksimalne vjerodostojnosti $\hat{\beta}_j$, dok vandijagonalni element na mjestu (i, j) označavamo s $\mathbf{Cov}(\beta_i, \beta_j)$ što je kovarijanca između procjenitelja maksimalne vjerodostojnosti $\hat{\beta}_i$ i $\hat{\beta}_j$. Procjenitelj varijance i kovarijance, u oznaci $\widehat{\mathbf{Var}}(\hat{\beta})$ dobivamo kao $\mathbf{Var}(\beta)$ za $\beta = \hat{\beta}$. Elemente te matrice označavamo s $\widehat{\mathbf{Var}}(\hat{\beta}_j)$ i $\widehat{\mathbf{Cov}}(\hat{\beta}_i, \hat{\beta}_j)$. Koristiti ćemo se i procjenom standardne pogreške procijenjenih parametara koja je oblika

$$\widehat{\mathbf{SE}}(\hat{\beta}_j) = (\widehat{\mathbf{Var}}(\hat{\beta}_j)), \ j = 0, 1, \ldots, p. \tag{3.9}$$

Koristimo se i formulacijom matrice informacije $\hat{\mathbb{I}}(\hat{\beta}) = \mathbb{X}^T \cdot \hat{V} \cdot \mathbb{X}$, gdje je $\mathbb{X}$ matrica oblika $n \cdot (p + 1)$, a $\hat{V}$ matrica oblika $n \cdot n$, te vrijedi

$$\mathbb{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix} \tag{3.10}$$

i

$$\hat{V} = \begin{bmatrix} \hat{\pi}(x_1) \cdot (1 - \hat{\pi}(x_1) & 0 & \ldots & 0 \\ 0 & \hat{\pi}(x_2) \cdot (1 - \hat{\pi}(x_2) & \ldots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \ldots & 0 & \hat{\pi}(x_n) \cdot (1 - \hat{\pi}(x_n) \end{bmatrix} \tag{3.11}$$

3.2 Testiranje značajnosti koeficijenata

Testiramo hipotezu da su svi parametri modela osim konstante β_0 jednaki nula, odnosno imamo

$$H_0 : \ \beta_1 = \cdots = \beta_p = 0. \tag{3.12}$$

Provodimo test omjera vjerodostojnosti i testna statistika G je oblika

$$G = -2 \cdot \ln\left(\frac{\left(\frac{n_1}{n}^{n_1}\right) \cdot \left(\frac{n_o}{n}^{n_0}\right)}{\prod_{i=1}^{n} \hat{\pi}(x_i)^{y_i} \cdot (1 - \hat{\pi}(x_i))^{1-y_i}} \right), \tag{3.13}$$

odnosno

$$G = -2 \cdot \left(\sum_{i=1}^{n} \left(y_i \cdot \ln\left(\hat{\pi}(x_i)\right) - (1-y_i) \cdot \ln\left(1-\hat{\pi}(x_i)\right) \right) - \left(n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n) \right) \right) \quad (3.14)$$

Pod hipotezom H_0 testna statistika G ima χ^2 razdiobu s p stupnjeva slobode, odnosno vrijedi

$$G \overset{H_0}{\sim} \chi^2(p). \quad (3.15)$$

Ukoliko odbacujemo nultu hipotezu, dobivamo da je barem jedan parametar različit od nule, stoga je korisno napraviti i univarijatni Waldov test za pojedine parametre, koji uz danu nultu hipotezu ima standardnu normalnu razdiobu.

Napomena 3.2.1. *Waldova testna statistika za hipotezu* $H_0 : \beta_j = 0$ *je oblika*

$$W_j = \frac{\hat{\beta}_j}{\widehat{\mathbf{SE}}(\hat{\beta}_j)} \quad (3.16)$$

Multivarijatni analogon Waldovog testa ima testnu statistiku oblika

$$W = \hat{\beta}^T \cdot (\widehat{\mathbf{Var}}(\hat{\beta}))^{-1} \cdot \hat{\beta} = \hat{\beta}^T \cdot (\mathbb{X}^T \hat{V} \mathbb{X}) \cdot \hat{\beta}, \quad (3.17)$$

te uz nultu hipotezu jednakosti svih parametara nuli ($H_0 : \beta_0 = \beta_1 = \cdots = \beta_p = 0$) ima χ^2 distribuciju s $p + 1$ stupnjeva slobode.

Napomena 3.2.2. *Intervali pouzdanosti dobivaju se jednako kao i u univarijatnom modelu, te dobivamo da su rubne točke* $(1 - \alpha) \cdot 100$ *% intervala pouzdanosti oblika*

$$\hat{\beta}_j \pm z_{1-\alpha/2} \cdot \widehat{\mathbf{SE}}(\beta_j), \ j = 0, 1, \ldots, p. \quad (3.18)$$

Poglavlje 4

Primjena logističke regresije

4.1 Pregled modela

Promatramo uzorak studenata upisanih od 2005. do 2011. godine. Zbog razlike u bodovanju bodova iz škole i zamjenom prijemnog ispita državnom maturom razdvajamo uzorak na dio prije i poslije državne mature. Također promatramo i dio svakog uzorka samo sa studentima koji su uspješno položili sve ispite iz prvog semestra studiranja. Za svaku nezavisnu varijablu i svaki dio uzorka provodimo univarijatnu logističku regresiju. Također, za svaki dio uzorka provodimo i multivarijatnu logističku regresiju s obje nezavisne varijable. Zavisna varijabla je uspjeh pojedinog studenta. Zavisna varijabla je dihotomna i označavamo ju s $\mathcal{US}$. Nezavisne varijable označavamo s $\mathcal{BS}$ (bodovi iz srednje škole) i $\mathcal{BT}$ (bodovi prijemnog ispita/državne mature). U tablici 4.1 dajemo jasniji pregled modela koje radimo. Oznaka ($\triangle$) pokraj nezavisne varijable označava da se radi o modelu na uzorku koji promatra samo studente koji su uspješno položili sve kolegije prvog semestra.

	Univarijatni model	Multivarijatni model
2005. - 2009.	$\mathcal{BS}$	$\mathcal{BS}$ + $\mathcal{BT}$
	$\mathcal{BT}$	
	$\mathcal{BS}$ ($\triangle$)	$\mathcal{BS}(\triangle)$ + $\mathcal{BT}(\triangle)$
	$\mathcal{BT}$ ($\triangle$)	
2010. - 2011.	$\mathcal{BS}$	$\mathcal{BS}$ + $\mathcal{BT}$
	$\mathcal{BT}$	
	$\mathcal{BS}$ ($\triangle$)	$\mathcal{BS}(\triangle)$ + $\mathcal{BT}(\triangle)$
	$\mathcal{BT}$ ($\triangle$)	

Tablica 4.1: Pregled modela

Prvo radimo univarijatne modele s nezavisnom varijablom $\mathcal{BS}$, zatim univarijatne modele s nezavisnom varijablom $\mathcal{BT}$. Na kraju poglavlja radimo multivarijatne modele i dajemo kratak zaključak i usporedbu promatranih modela na temelju pripadnih ROC krivulja i AUC površina ispod krivulja.

4.2 Univarijatni modeli: Bodovi iz škole

Uzorak: 2005. - 2009. godina

Promatramo dio uzorka prije uvođenja državne mature, odnosno promatramo studente koji su upisali studij između 2005. i 2009. godine uključeno. Za svakog studenta promatramo broj ostvarenih bodova iz srednje škole kao nezavisnu varijablu, te uspjeh na prvoj godini studije kao zavisnu, dihotomnu varijablu. Varijablu uspjeh studenta označavamo s $\mathcal{US}$, dok nezavisnu varijablu označavamo s $\mathcal{BS}$. Tablica 4.2 daje osnovne podatke o promatranim varijablama, po promatranim godinama.

g	n_g	raspon $\mathcal{BS}$	$\mathcal{US}$
2005	227	0 – 260	0 ili 1
2006	223	0 – 260	0 ili 1
2007	179	0 – 260	0 ili 1
2008	181	0 – 260	0 ili 1
2009	196	0 – 260	0 ili 1
ukupno	1006	0 – 260	0 ili 1

Tablica 4.2: Osnovni podaci za promatrani uzorak

Provodimo logističku regresiju za dani uzorak od $n = 1006$ studenata. Parametre modela procjenjujemo metodom maksimalne vjerodostojnosti. Procijenjeni parametri, kao i procjena standardne pogreške parametara i 95 % pouzdani intervali za procijenjene parametre dani su u tablici 4.3.

	parametar	procjena parametra	$\mathbf{SE}(\hat{\beta}_i)$	95 % interval pouzdanosti
konstanta	$\hat{\beta}_0$	-12.099677	1.003077	$\langle -14.14056429, -10.20635587 \rangle$
$\mathcal{BS}$	$\hat{\beta}_1$	0.047727	0.004044	$\langle 0.04007447, 0.05593385 \rangle$

Tablica 4.3: Procjena parametara

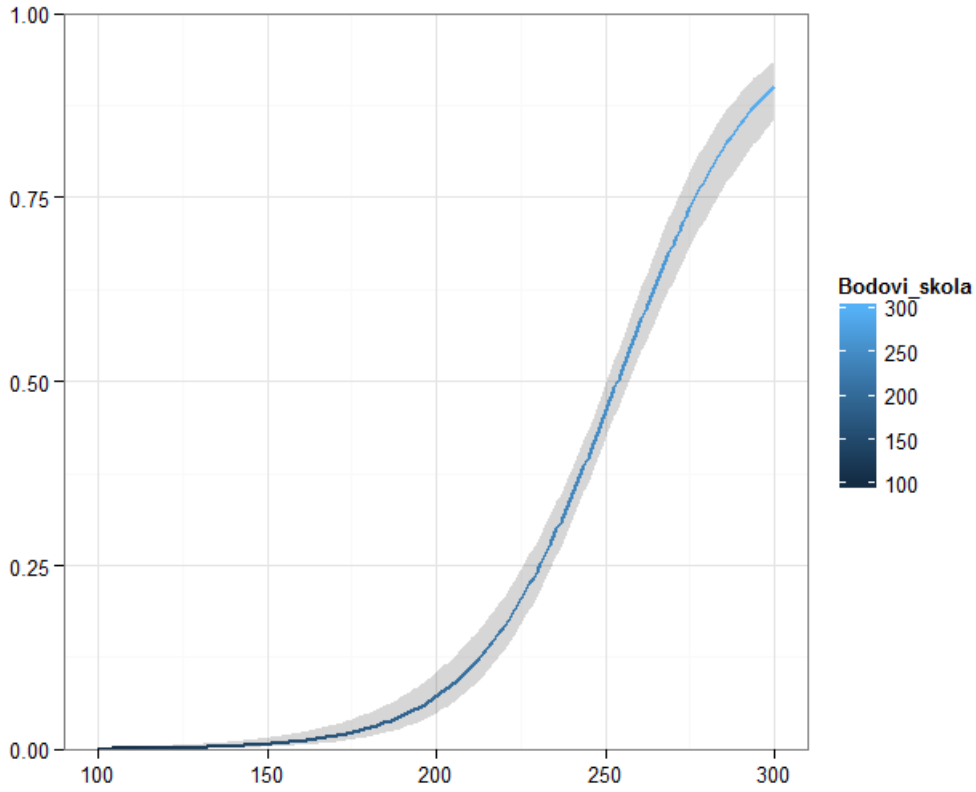
Dobivamo da je procijenjena logaritamska transformacija uvjetnog očekivanja oblika

$$\hat{g}(\mathcal{BS}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathcal{BS} = -12.099677 + 0.047727 \cdot \mathcal{BS}, \tag{4.1}$$

odnosno da je procjena uvjetnog očekivanja od dihotomne varijable $\mathcal{US}$ uz dano $\mathcal{BS} = x$ oblika

$$\hat{\pi}(x) = \frac{e^{\hat{\beta_0}+\hat{\beta_1}x}}{1+e^{\hat{\beta_0}+\hat{\beta_1}x}} = \frac{e^{-12.099677+0.047727x}}{1+e^{-12.099677+0.047727x}} \tag{4.2}$$

Slika 4.1 daje grafički prikaz "S-krivulje" danog modela uz pripadne 95 %-tne intervale. Interpretacija grafičkog prikaza je da za danog studenta koji je upisao studij matematika s brojem bodova iz škole jednak $\mathcal{BS} = x$, $\hat{\pi}(x)$ predstavlja procjenu vjerojatnosti da student uspješno završi prvu godinu studija.



Slika 4.1: Grafički prikaz procijenjene logističke funkcije

Testirajmo sada značajnost procijenjenih parametara. Testiramo hipoteze

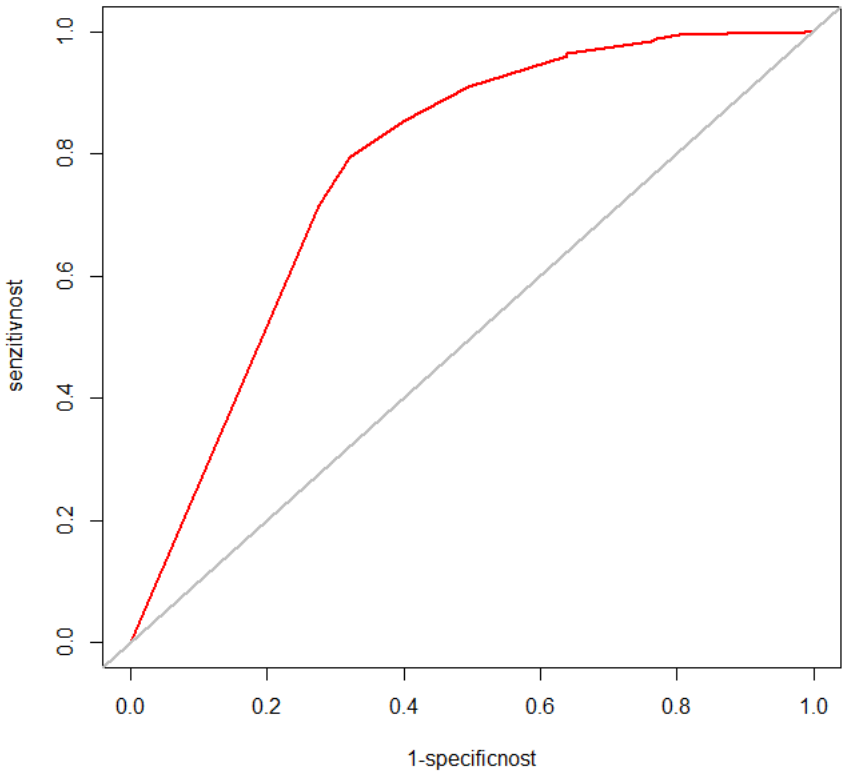
$$\beta_i = 0,\ i = 0,\ 1. \tag{4.3}$$

Provodimo Waldov test koji je opisan u teorijskom djelu. Testnu statistiku Waldovog testa označavamo s $\mathbf{W}$ ($\mathbf{W} \overset{H_0}{\sim} N(0,1)$). Rezultate testa dajemo u tablici 4.4

H_0 :	**W**	$p-$vrijednost
$\beta_0 = 0$	-12.06	< 0.0001
$\beta_1 = 0$	11.80	< 0.0001

Tablica 4.4: Testovi značajnosti parametara modela

Iz tablice 4.4 vidimo da na svakom razumnom nivou značajnosti odbacujemo nulte hipoteze, odnosno da su promatrani parametri značajni za model. Na slici 4.2 dajemo ROC krivulju. Površina ispod ROC krivulje, odnosno AUC promatranog modela iznosi 0.7728.



Slika 4.2: ROC krivulja

Promotrimo sada samo studente koji su uspješno položili sve kolegije prvog semestra studija. Provodimo logističku regresiju za dani uzorak od $n = 484$ studenata. Parame-

tre modela procjenjujemo metodom maksimalne vjerodostojnosti. Procijenjeni parametri, kao i procjena standardne pogreške parametara i 95 % pouzdani intervali za procijenjene parametre dani su u tablici 4.5.

	parametar	procjena parametra	$\mathbf{SE}(\hat{\beta}_i)$	95 % interval pouzdanosti
konstanta	$\hat{\beta}_0$	-7.277622	1.270452	$\langle -9.85062656, -4.87128191 \rangle$
$\mathcal{BS}$	$\hat{\beta}_1$	0.033762	0.005174	$\langle 0.02396256, 0.04424114 \rangle$

Tablica 4.5: Procjena parametara

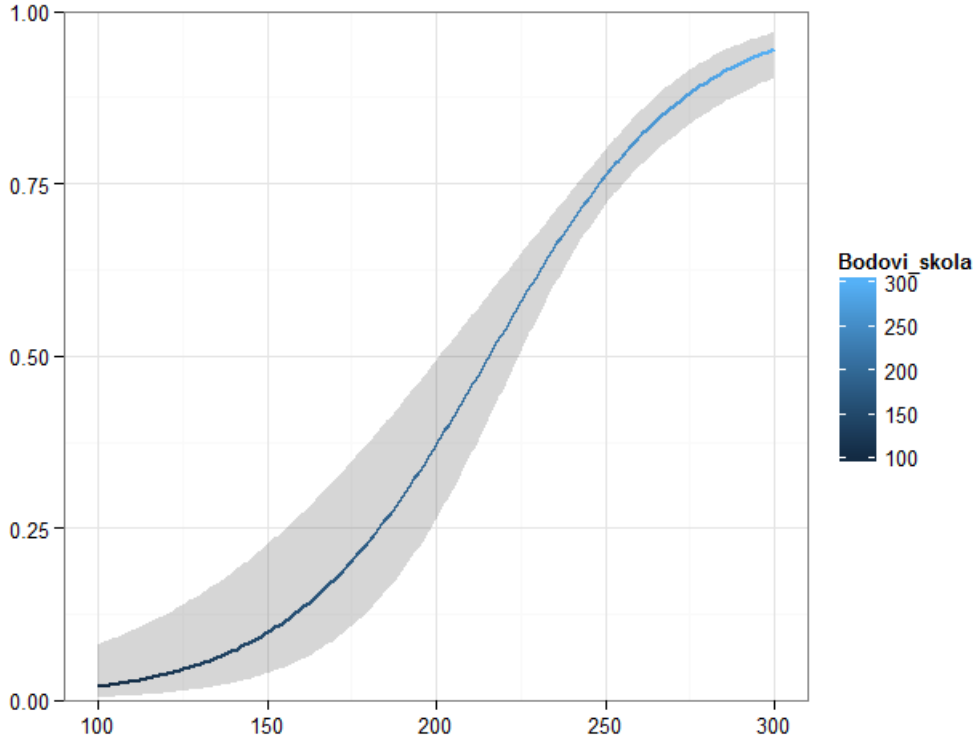
Dobivamo da je procijenjena logaritamska transformacija uvjetnog očekivanja oblika

$$\hat{g}(\mathcal{BS}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathcal{BS} = -7.277622 + 0.033762 \cdot \mathcal{BS}, \tag{4.4}$$

odnosno da je procjena uvjetnog očekivanja od dihotomne varijable $\mathcal{US}$ uz dano $\mathcal{BS} = x$ oblika

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}} = \frac{e^{-7.277622 + 0.033762 x}}{1 + e^{-7.277622 + 0.033762 x}} \tag{4.5}$$

Slika 4.3 daje grafički prikaz "S-krivulje" danog modela uz pripadne 95 %-tne intervale. Interpretacija grafičkog prikaza je da za danog studenta koji je upisao studij matematika s brojem bodova iz škole jednak $\mathcal{BS} = x$, $\hat{\pi}(x)$ predstavlja procjenu vjerojatnosti da student uspješno završi prvu godinu studija.



Slika 4.3: Grafički prikaz procijenjene logističke funkcije

Testirajmo sada značajnost procijenjenih parametara. Testiramo hipoteze

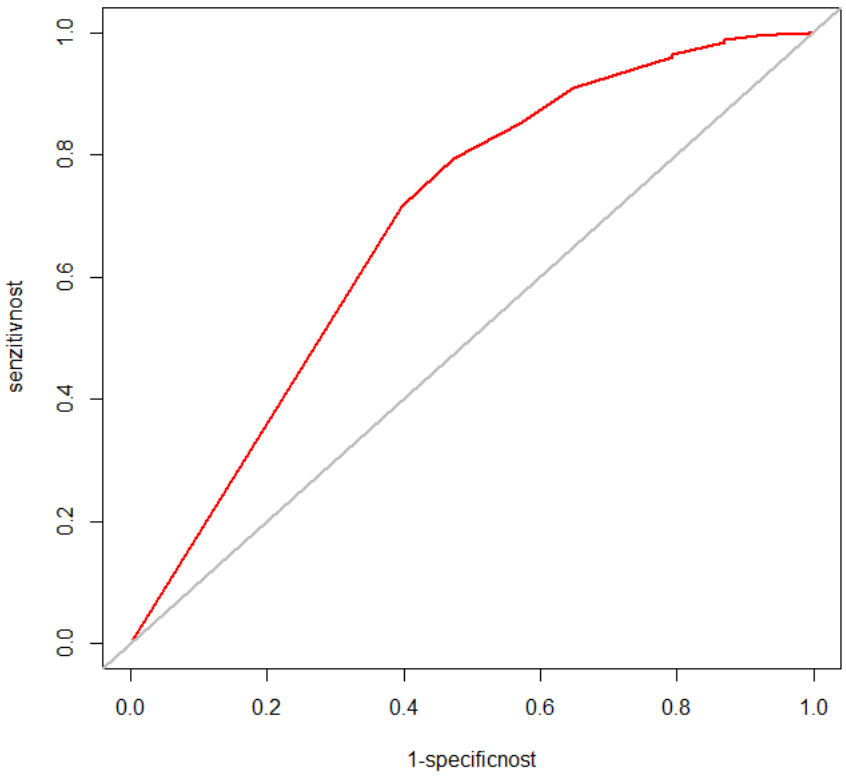
$$\beta_i = 0,\ i = 0,\ 1. \tag{4.6}$$

Provodimo Waldov test koji je opisan u teorijskom djelu. Testnu statistiku Waldovog testa označavamo s $\mathbf{W}$ ($\mathbf{W} \overset{H_0}{\sim} N(0,1)$). Rezultate testa dajemo u tablici 4.6

H_0 :	$\mathbf{W}$	$p-$vrijednost
$\beta_0 = 0$	-5.728	< 0.0001
$\beta_1 = 0$	6.525	< 0.0001

Tablica 4.6: Testovi značajnosti parametara modela

Iz tablice 4.6 vidimo da na svakom razumnom novou značajnosti odbacujemo nulte hipoteze, odnosno da su promatrani parametri značajni za model. Na slici 4.4 dajemo ROC krivulju. Površina ispod ROC krivulje, odnosno AUC promatranog modela iznosi 0.6871.



Slika 4.4: ROC krivulja

Uzorak: 2010., 2011. godina

Promatramo dio uzorka nakon uvođenja državne mature, odnosno promatramo studente koji su upisali studij 2010. i 2011. godine. Za svakog studenta promatramo broj ostvarenih bodova iz srednje škole kao nezavisnu varijablu, te uspjeh na prvoj godini studije kao zavisnu, dihotomnu varijablu. Varijablu uspjeh studenta označavamo s $\mathcal{US}$, dok nezavisnu varijablu označavamo s $\mathcal{BS}$. Tablica 4.7 daje osnovne podatke o promatranim varijablama, po promatranim godinama.

g	n_g	raspon $\mathcal{BS}$	$\mathcal{US}$
2010	198	0 – 300	0 ili 1
2011	197	0 – 300	0 ili 1
ukupno	395	0 – 300	0 ili 1

Tablica 4.7: Osnovni podaci za promatrani uzorak

Provodimo logističku regresiju za dani uzorak od $n = 395$ studenata. Parametre modela procjenjujemo metodom maksimalne vjerodostojnosti. Procijenjeni parametri, kao i procjena standardne pogreške parametara i 95 % pouzdani intervali za procijenjene parametre dani su u tablici 4.8.

	parametar	procjena parametra	$\mathbf{SE}(\hat{\beta}_i)$	95 % interval pouzdanosti
konstanta	$\hat{\beta}_0$	-11.288787	1.614588	$\langle -14.6123171, -8.27089797 \rangle$
$\mathcal{BS}$	$\hat{\beta}_1$	0.039228	0.005819	$\langle 0.0283163, 0.05116864 \rangle$

Tablica 4.8: Procjena parametara

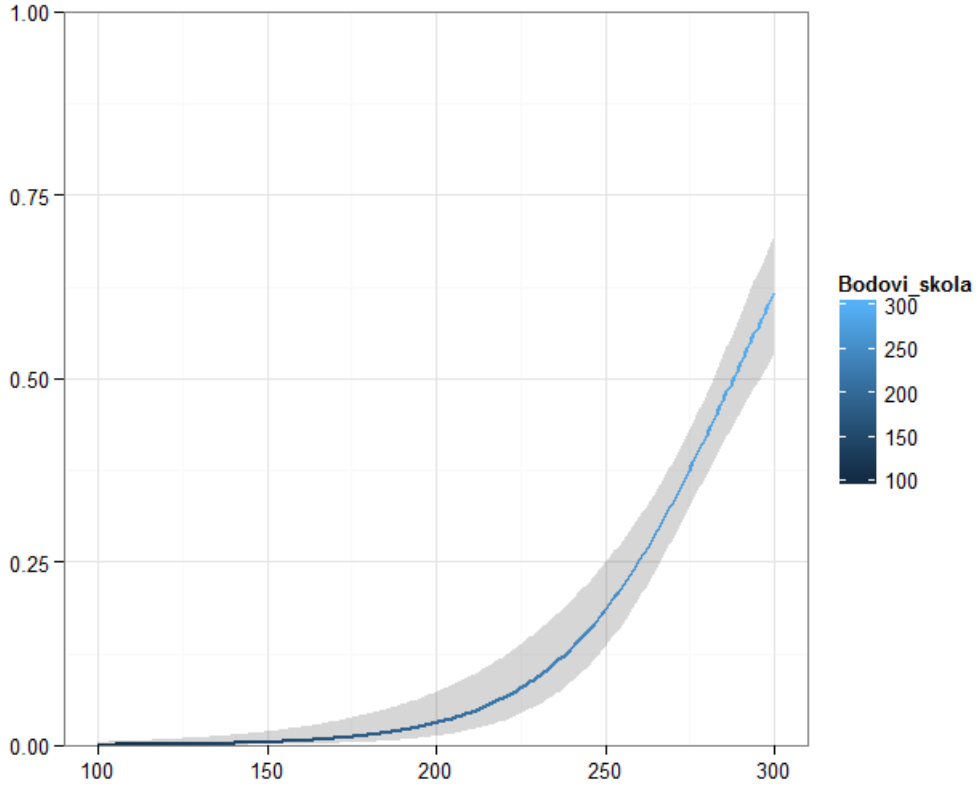
Dobivamo da je procijenjena logaritamska transformacija uvjetnog očekivanja oblika

$$\hat{g}(\mathcal{BS}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathcal{BS} = -11.288787 + 0.039228 \cdot \mathcal{BS}, \tag{4.7}$$

odnosno da je procjena uvjetnog očekivanja od dihotomne varijable $\mathcal{US}$ uz dano $\mathcal{BS} = x$ oblika

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}} = \frac{e^{-11.288787 + 0.039228 x}}{1 + e^{-11.288787 + 0.039228 x}} \tag{4.8}$$

Slika 4.5 daje grafički prikaz "S-krivulje" danog modela uz pripadne 95 %-tne intervale. Interpretacija grafičkog prikaza je da za danog studenta koji je upisao studij matematika s brojem bodova iz škole jednak $\mathcal{BS} = x$, $\hat{\pi}(x)$ predstavlja procjenu vjerojatnosti da student uspješno završi prvu godinu studija.



Slika 4.5: Grafički prikaz procijenjene logističke funkcije

Testirajmo sada značajnost procijenjenih parametara. Testiramo hipoteze

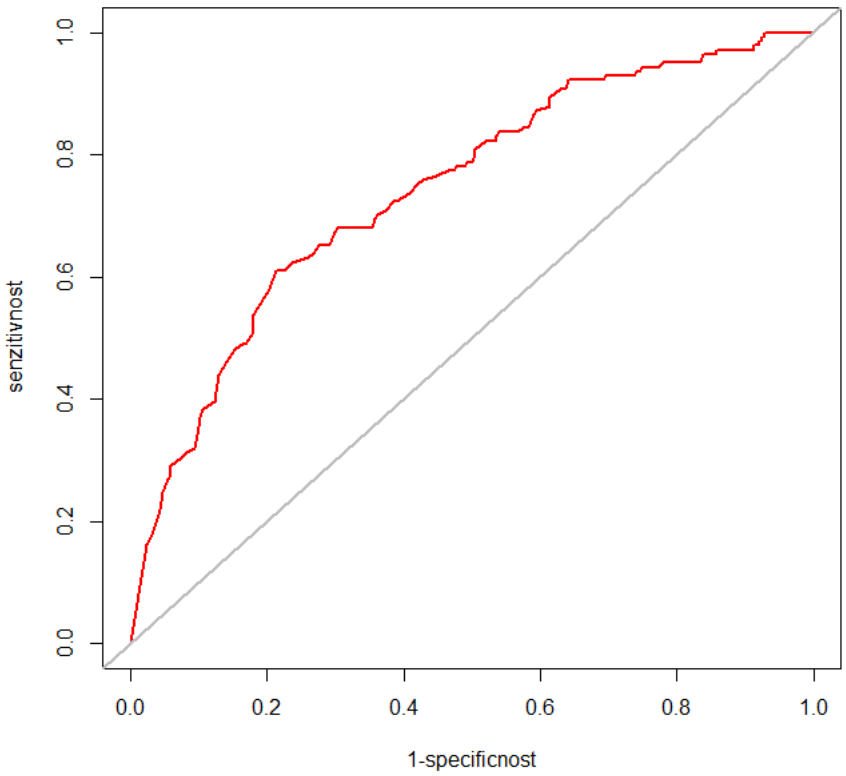
$$\beta_i = 0,\ i = 0,\ 1. \tag{4.9}$$

Provodimo Waldov test koji je opisan u teorijskom djelu. Testnu statistiku Waldovog testa označavamo s $\mathbf{W}$ ($\mathbf{W} \overset{H_0}{\sim} N(0,1)$). Rezultate testa dajemo u tablici 4.9

H_0 :	$\mathbf{W}$	$p-$vrijednost
$\beta_0 = 0$	-6.992	< 0.0001
$\beta_1 = 0$	6.742	< 0.0001

Tablica 4.9: Testovi značajnosti parametara modela

Iz tablice 4.9 vidimo da na svakom razumnom nivou značajnosti odbacujemo nulte hipoteze, odnosno da su promatrani parametri značajni za model. Na slici 4.6 dajemo ROC krivulju. Površina ispod ROC krivulje, odnosno AUC promatranog modela iznosi 0.742.



Slika 4.6: ROC krivulja

Promotrimo sada samo studente koji su uspješno položili sve kolegije prvog semestra studija. Provodimo logističku regresiju za dani uzorak od $n = 169$ studenata. Parametre modela procjenjujemo metodom maksimalne vjerodostojnosti. Procijenjeni parametri, kao i procjena standardne pogreške parametara i 95 % pouzdani intervali za procijenjene parametre dani su u tablici 4.10.

	parametar	procjena parametra	$\mathbf{SE}(\hat{\beta}_i)$	95 % interval pouzdanosti
konstanta	$\hat{\beta}_0$	-3.437716	2.337239	$\langle -7.999316861, 1.28813706 \rangle$
$\mathcal{BS}$	$\hat{\beta}_1$	0.018256	0.008493	$\langle 0.001234916, 0.03494497 \rangle$

Tablica 4.10: Procjena parametara

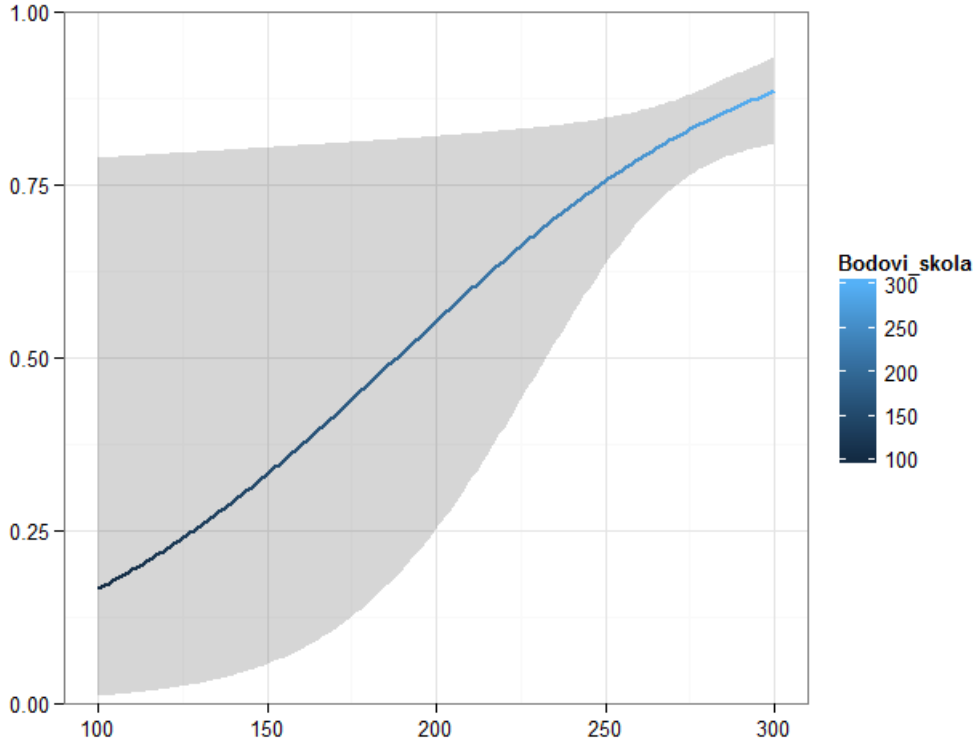
Dobivamo da je procijenjena logaritamska transformacija uvjetnog očekivanja oblika

$$\hat{g}(\mathcal{BS}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathcal{BS} = -3.437716 + 0.018256 \cdot \mathcal{BS}, \tag{4.10}$$

odnosno da je procjena uvjetnog očekivanja od dihotomne varijable $\mathcal{US}$ uz dano $\mathcal{BS} = x$ oblika

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}} = \frac{e^{-3.437716 + 0.018256 x}}{1 + e^{-3.437716 + 0.018256 x}} \tag{4.11}$$

Slika 4.7 daje grafički prikaz "S-krivulje" danog modela uz pripadne 95 %-tne intervale. Interpretacija grafičkog prikaza je da za danog studenta koji je upisao studij matematika s brojem bodova iz škole jednak $\mathcal{BS} = x$, $\hat{\pi}(x)$ predstavlja procjenu vjerojatnosti da student uspješno završi prvu godinu studija.



Slika 4.7: Grafički prikaz procijenjene logističke funkcije

Testirajmo sada značajnost procijenjenih parametara. Testiramo hipoteze

$$\beta_i = 0,\ i = 0,\ 1. \tag{4.12}$$

Provodimo Waldov test koji je opisan u teorijskom djelu. Testnu statistiku Waldovog testa označavamo s $\mathbf{W}$ ($\mathbf{W} \overset{H_0}{\sim} N(0,1)$). Rezultate testa dajemo u tablici 4.6

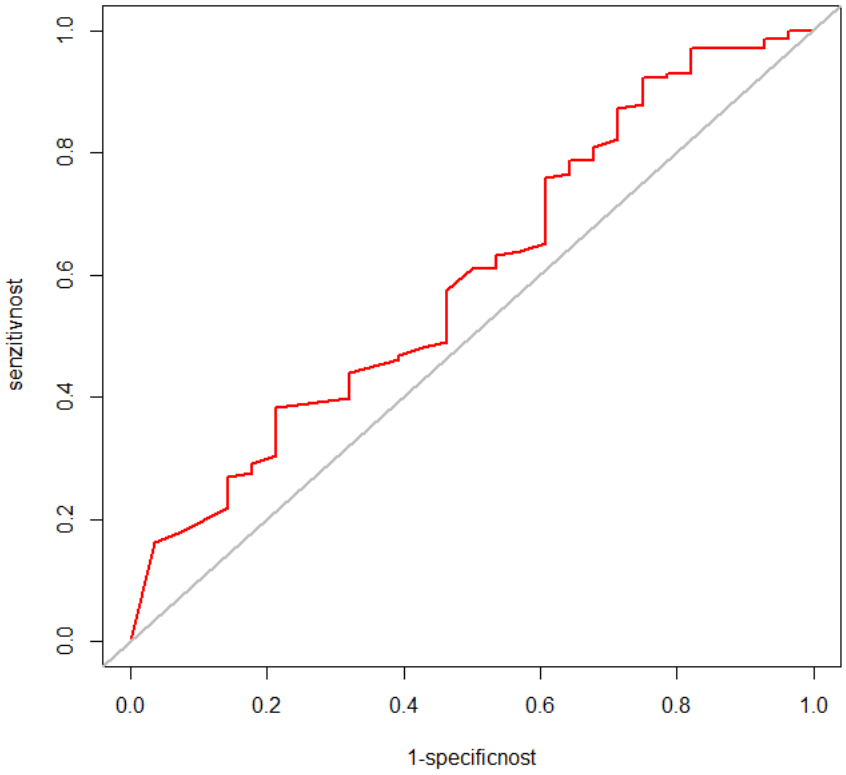
H_0 :	$\mathbf{W}$	$p-$vrijednost
$\beta_0 = 0$	-1.471	0.1413
$\beta_1 = 0$	2.150	0.0316

Tablica 4.11: Testovi značajnosti parametara modela

Iz tablice 4.11 vidimo da hipotezu

$$H_0 : \beta_0 = 0 \tag{4.13}$$

ne možemo odbaciti, dok hipotezu da je $\beta_1 = 0$ odbacujemo na nivou značajnosti $\alpha = 0.05$. Na slici 4.8 dajemo ROC krivulju. Površina ispod ROC krivulje, odnosno AUC promatranog modela iznosi 0.5974



Slika 4.8: ROC krivulja

4.3 Univarijatni modeli: Bodovi prijemnog ispita/državne mature

Uzorak: 2005. - 2009. godina

Promatramo dio uzorka prije uvođenja državne mature, odnosno promatramo studente koji su upisali studij između 2005. i 2009. godine uključeno. Za svakog studenta promatramo broj ostvarenih bodova na prijemnom ispitu kao nezavisnu varijablu, te uspjeh na prvoj godini studije kao zavisnu, dihotomnu varijablu. Varijablu uspjeh studenta označavamo s $\mathcal{US}$, dok nezavisnu varijablu označavamo s $\mathcal{BT}$. Tablica 4.12 daje osnovne podatke o promatranim varijablama, po promatranim godinama.

g	n_g	raspon $\mathcal{BT}$	$\mathcal{US}$
2005	227	0 – 663	0 ili 1
2006	223	0 – 663	0 ili 1
2007	179	0 – 663	0 ili 1
2008	181	0 – 663	0 ili 1
2009	196	0 – 663	0 ili 1
ukupno	1006	0 – 663	0 ili 1

Tablica 4.12: Osnovni podaci za promatrani uzorak

Provodimo logističku regresiju za dani uzorak od $n = 1006$ studenata. Parametre modela procjenjujemo metodom maksimalne vjerodostojnosti. Procijenjeni parametri, kao i procjena standardne pogreške parametara i 95 % pouzdani intervali za procijenjene parametre dani su u tablici 4.13.

	parametar	procjena parametra	$\mathbf{SE}(\hat{\beta}_i)$	95 % interval pouzdanosti
konstanta	$\hat{\beta}_0$	-3.431812	0.214798	$\langle -3.865391719, -3.022445498 \rangle$
$\mathcal{BT}$	$\hat{\beta}_1$	0.008698	0.000623	$\langle 0.007516259, 0.009961378 \rangle$

Tablica 4.13: Procjena parametara

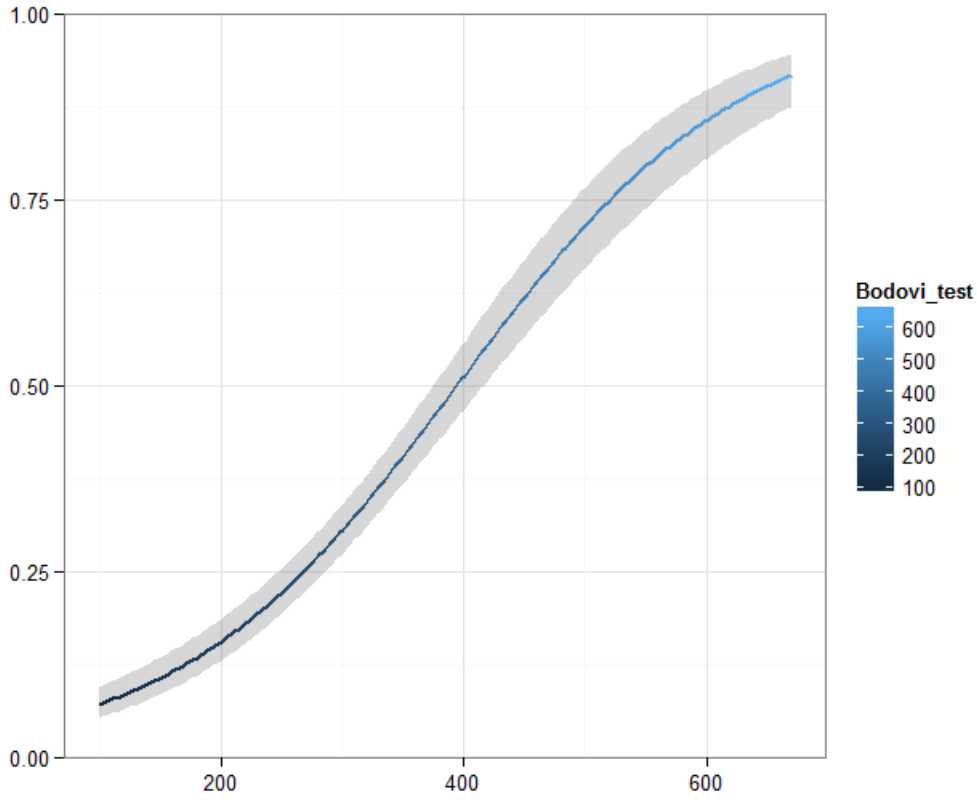
Dobivamo da je procijenjena logaritamska transformacija uvjetnog očekivanja oblika

$$\hat{g}(\mathcal{BT}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathcal{BT} = -3.431812 + 0.008698 \cdot \mathcal{BT}, \tag{4.14}$$

odnosno da je procjena uvjetnog očekivanja od dihotomne varijable $\mathcal{US}$ uz dano $\mathcal{BT} = x$ oblika

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x}} = \frac{e^{-3.431812+0.008698x}}{1 + e^{-3.431812+0.008698x}} \tag{4.15}$$

Slika 4.9 daje grafički prikaz "S-krivulje" danog modela uz pripadne 95 %-tne intervale. Interpretacija grafičkog prikaza je da za danog studenta koji je upisao studij matematika s brojem bodova iz škole jednak $\mathcal{BT} = x$, $\hat{\pi}(x)$ predstavlja procjenu vjerojatnosti da student uspješno završi prvu godinu studija.



Slika 4.9: Grafički prikaz procijenjene logističke funkcije

Testirajmo sada značajnost procijenjenih parametara. Testiramo hipoteze

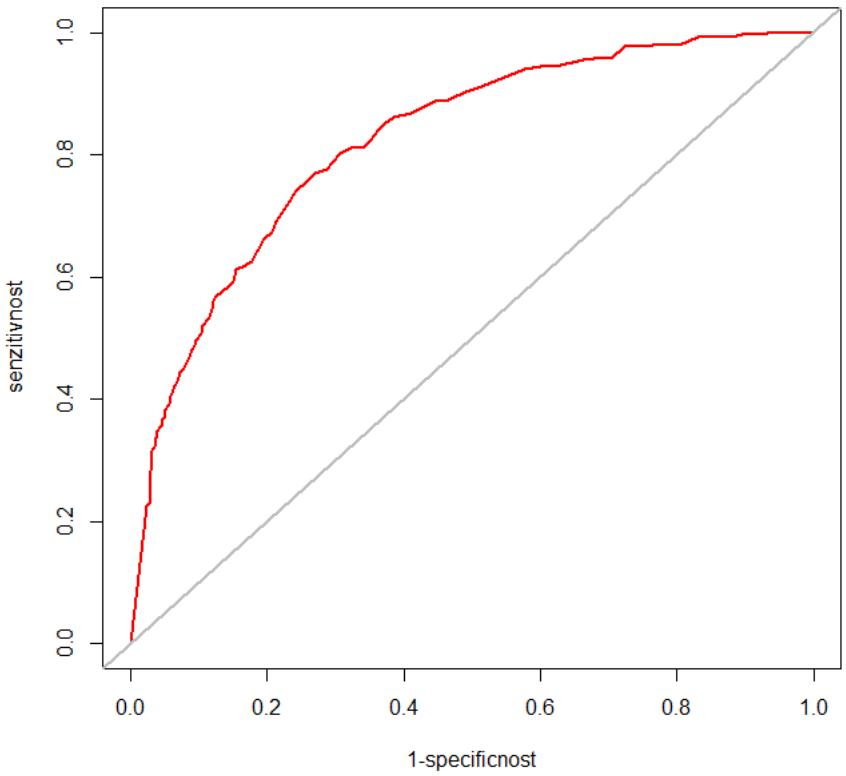
$$\beta_i = 0,\ i = 0,\ 1. \tag{4.16}$$

Provodimo Waldov test koji je opisan u teorijskom djelu. Testnu statistiku Waldovog testa označavamo s **W** ($\mathbf{W} \overset{H_0}{\sim} N(0,1)$). Rezultate testa dajemo u tablici 4.14

H_0 :	**W**	$p-$vrijednost
$\beta_0 = 0$	-15.98	< 0.0001
$\beta_1 = 0$	13.96	< 0.0001

Tablica 4.14: Testovi značajnosti parametara modela

Iz tablice 4.14 vidimo da na svakom razumnom nivou značajnosti odbacujemo nulte hipoteze, odnosno da su promatrani parametri značajni za model. Na slici 4.10 dajemo ROC krivulju. Površina ispod ROC krivulje, odnosno AUC promatranog modela iznosi 0.822.



Slika 4.10: ROC krivulja

Promotrimo sada samo studente koji su uspješno položili sve kolegije prvog semestra studija. Provodimo logističku regresiju za dani uzorak od $n = 484$ studenata. Parametre modela procjenjujemo metodom maksimalne vjerodostojnosti. Procijenjeni parametri, kao i procjena standardne pogreške parametara i 95 % pouzdani intervali za procijenjene parametre dani su u tablici 4.15.

	parametar	procjena parametra	$\mathbf{SE}(\hat{\beta}_i)$	95 % interval pouzdanosti
konstanta	$\hat{\beta}_0$	-1.3548767	0.3163771	$\langle -1.991895605, -0.748771795 \rangle$
$\mathcal{BT}$	$\hat{\beta}_1$	0.0066651	0.0009265	$\langle 0.0049297350.008570345 \rangle$

Tablica 4.15: Procjena parametara

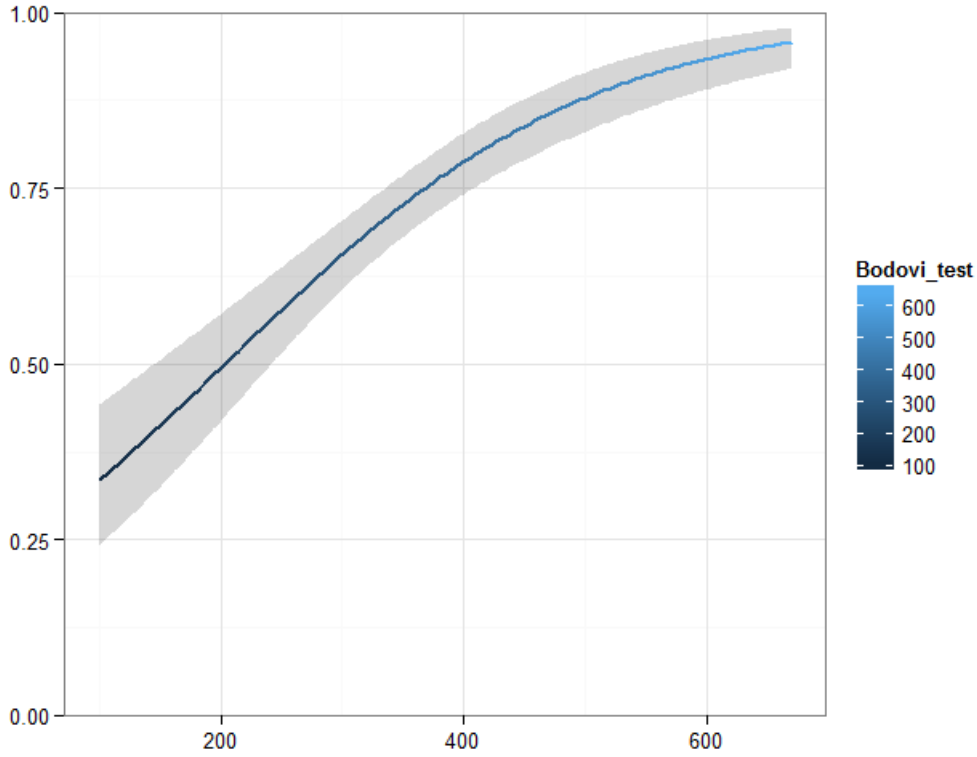
Dobivamo da je procijenjena logaritamska transformacija uvjetnog očekivanja oblika

$$\hat{g}(\mathcal{BT}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathcal{BT} = -1.3548767 + 0.0066651 \cdot \mathcal{BT}, \tag{4.17}$$

odnosno da je procjena uvjetnog očekivanja od dihotomne varijable $\mathcal{US}$ uz dano $\mathcal{BT} = x$ oblika

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x}} = \frac{e^{-1.3548767+0.0066651x}}{1 + e^{-1.3548767+0.0066651x}} \tag{4.18}$$

Slika 4.11 daje grafički prikaz "S-krivulje" danog modela uz pripadne 95 %-tne intervale. Interpretacija grafičkog prikaza je da za danog studenta koji je upisao studij matematika s brojem bodova iz škole jednak $\mathcal{BT} = x$, $\hat{\pi}(x)$ predstavlja procjenu vjerojatnosti da student uspješno završi prvu godinu studija.



Slika 4.11: Grafički prikaz procijenjene logističke funkcije

Testirajmo sada značajnost procijenjenih parametara. Testiramo hipoteze

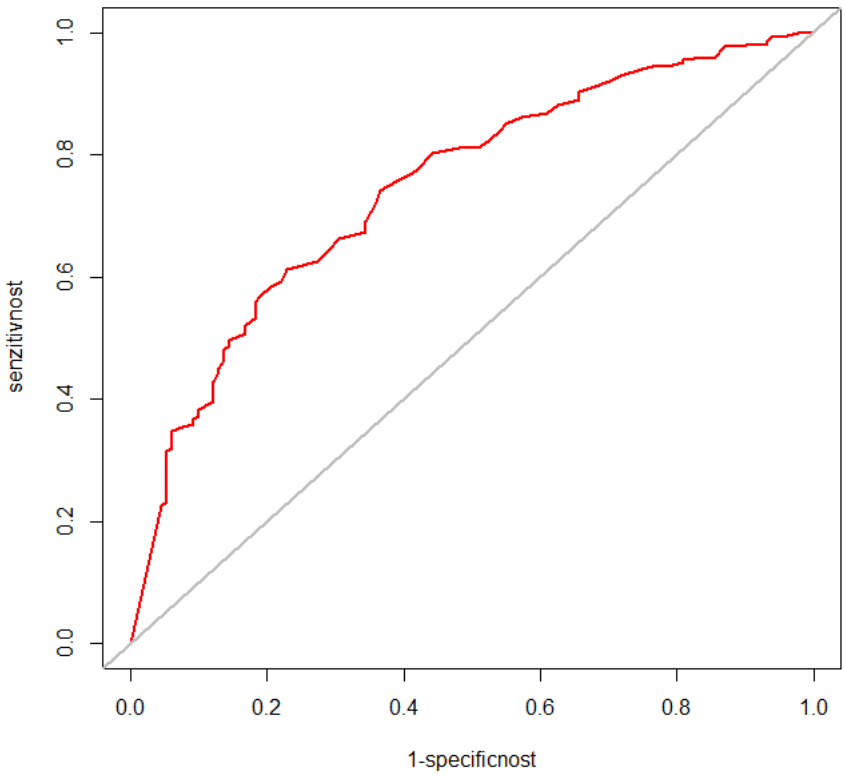
$$\beta_i = 0,\ i = 0,\ 1. \tag{4.19}$$

Provodimo Waldov test koji je opisan u teorijskom djelu. Testnu statistiku Waldovog testa označavamo s $\mathbf{W}$ ($\mathbf{W} \overset{H_0}{\sim} N(0,1)$). Rezultate testa dajemo u tablici 4.16

H_0 :	$\mathbf{W}$	$p-$vrijednost
$\beta_0 = 0$	-4.282	< 0.0001
$\beta_1 = 0$	7.194	< 0.0001

Tablica 4.16: Testovi značajnosti parametara modela

Iz tablice 4.16 vidimo da na svakom razumnom novou značajnosti odbacujemo nulte hipoteze, odnosno da su promatrani parametri značajni za model. Na slici 4.12 dajemo ROC krivulju. Površina ispod ROC krivulje, odnosno AUC promatranog modela iznosi 0.7462.



Slika 4.12: ROC krivulja

Uzorak: 2010., 2011. godina

Promatramo dio uzorka nakon uvođenja državne mature, odnosno promatramo studente koji su upisali studij 2010. i 2011. godine. Za svakog studenta promatramo broj ostvarenih bodova na državnoj maturi kao nezavisnu varijablu, te uspjeh na prvoj godini studije kao zavisnu, dihotomnu varijablu. Varijablu uspjeh studenta označavamo s $\mathcal{US}$, dok nezavisnu varijablu označavamo s $\mathcal{BT}$. Tablica 4.17 daje osnovne podatke o promatranim varijablama, po promatranim godinama.

g	n_g	raspon $\mathcal{BT}$	$\mathcal{US}$
2010	198	0 – 600	0 ili 1
2011	197	0 – 600	0 ili 1
ukupno	395	0 – 600	0 ili 1

Tablica 4.17: Osnovni podaci za promatrani uzorak

Provodimo logističku regresiju za dani uzorak od $n = 395$ studenata. Parametre modela procjenjujemo metodom maksimalne vjerodostojnosti. Procijenjeni parametri, kao i procjena standardne pogreške parametara i 95 % pouzdani intervali za procijenjene parametre dani su u tablici 4.18.

	parametar	procjena parametra	$\mathbf{SE}(\hat{\beta}_i)$	95 % interval pouzdanosti
konstanta	$\hat{\beta}_0$	-13.77820	1.45517	$\langle -16.80090884, -11.08308700 \rangle$
$\mathcal{BT}$	$\hat{\beta}_1$	0.02596	0.00278	$\langle 0.02079729, 0.03172114 \rangle$

Tablica 4.18: Procjena parametara

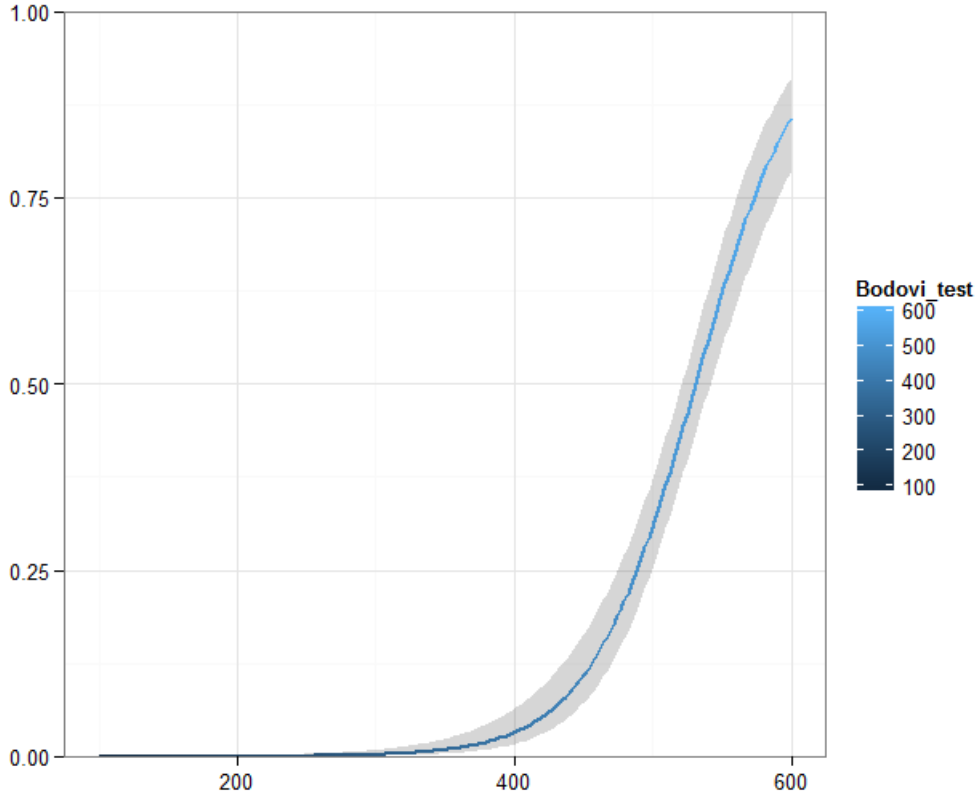
Dobivamo da je procijenjena logaritamska transformacija uvjetnog očekivanja oblika

$$\hat{g}(\mathcal{BT}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathcal{BT} = -13.77820 + 0.02596 \cdot \mathcal{BT}, \tag{4.20}$$

odnosno da je procjena uvjetnog očekivanja od dihotomne varijable $\mathcal{US}$ uz dano $\mathcal{BT} = x$ oblika

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}} = \frac{e^{-13.77820 + 0.02596 x}}{1 + e^{-13.77820 + 0.02596 x}} \tag{4.21}$$

Slika 4.13 daje grafički prikaz "S-krivulje" danog modela uz pripadne 95 %-tne intervale. Interpretacija grafičkog prikaza je da za danog studenta koji je upisao studij matematika s brojem bodova iz škole jednak $\mathcal{BT} = x$, $\hat{\pi}(x)$ predstavlja procjenu vjerojatnosti da student uspješno završi prvu godinu studija.



Slika 4.13: Grafički prikaz procijenjene logističke funkcije

Testirajmo sada značajnost procijenjenih parametara. Testiramo hipoteze

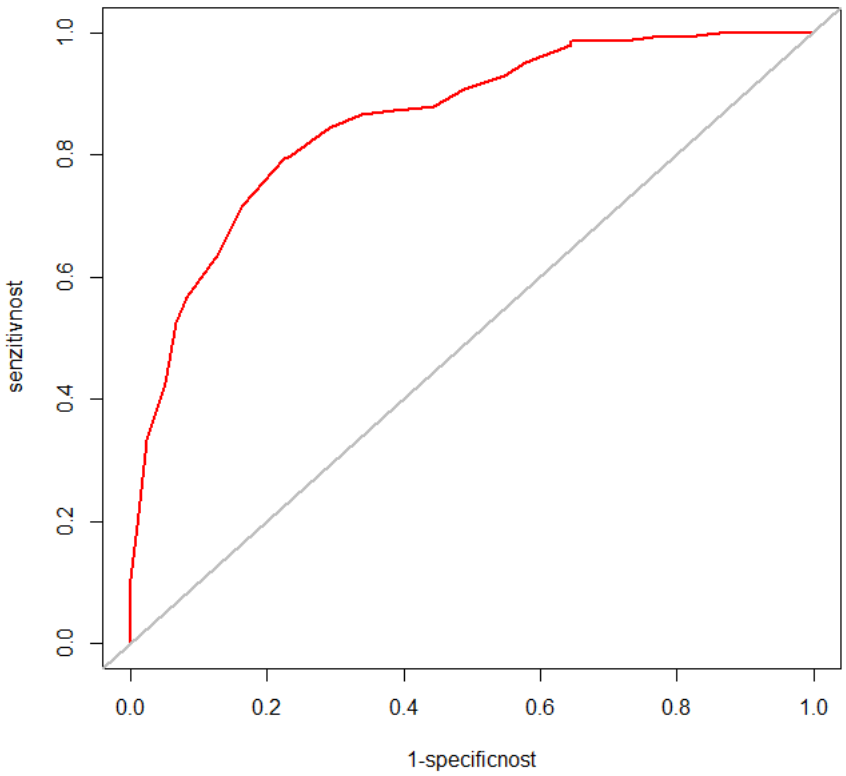
$$\beta_i = 0,\ i = 0,\ 1. \tag{4.22}$$

Provodimo Waldov test koji je opisan u teorijskom djelu. Testnu statistiku Waldovog testa označavamo s $\mathbf{W}$ ($\mathbf{W} \overset{H_0}{\sim} N(0,1)$). Rezultate testa dajemo u tablici 4.19

H_0 :	$\mathbf{W}$	$p-$vrijednost
$\beta_0 = 0$	-9.468	< 0.0001
$\beta_1 = 0$	9.338	< 0.0001

Tablica 4.19: Testovi značajnosti parametara modela

Iz tablice 4.19 vidimo da na svakom razumnom novou značajnosti odbacujemo nulte hipoteze, odnosno da su promatrani parametri značajni za model. Na slici 4.14 dajemo ROC krivulju. Površina ispod ROC krivulje, odnosno AUC promatranog modela iznosi 0.856.



Slika 4.14: ROC krivulja

Promotrimo sada samo studente koji su uspješno položili sve kolegije prvog semestra studija. Provodimo logističku regresiju za dani uzorak od $n = 169$ studenata. Parametre modela procjenjujemo metodom maksimalne vjerodostojnosti. Procijenjeni parametri, kao i procjena standardne pogreške parametara i 95 % pouzdani intervali za procijenjene parametre dani su u tablici 4.20.

	parametar	procjena parametra	$\mathbf{SE}(\hat{\beta}_i)$	95 % interval pouzdanosti
konstanta	$\hat{\beta}_0$	-6.846703	2.094648	$\langle -11.115573113, -2.8348359 \rangle$
$\mathcal{BT}$	$\hat{\beta}_1$	0.016145	0.004061	$\langle 0.008437824, 0.0244847 \rangle$

Tablica 4.20: Procjena parametara

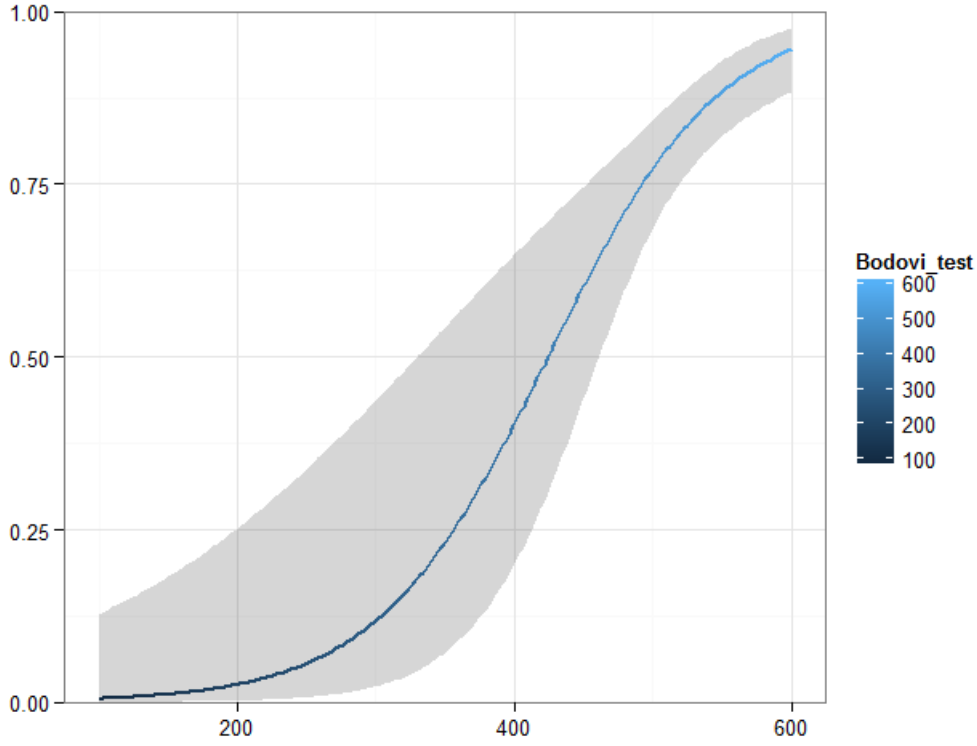
Dobivamo da je procijenjena logaritamska transformacija uvjetnog očekivanja oblika

$$\hat{g}(\mathcal{BT}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathcal{BT} = -6.846703 + 0.016145 \cdot \mathcal{BT}, \tag{4.23}$$

odnosno da je procjena uvjetnog očekivanja od dihotomne varijable $\mathcal{US}$ uz dano $\mathcal{BT} = x$ oblika

$$\hat{\pi}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}} = \frac{e^{-6.846703 + 0.016145 x}}{1 + e^{-6.846703 + 0.016145 x}} \tag{4.24}$$

Slika 4.15 daje grafički prikaz "S-krivulje" danog modela uz pripadne 95 %-tne intervale. Interpretacija grafičkog prikaza je da za danog studenta koji je upisao studij matematika s brojem bodova iz škole jednak $\mathcal{BT} = x$, $\hat{\pi}(x)$ predstavlja procjenu vjerojatnosti da student uspješno završi prvu godinu studija.



Slika 4.15: Grafički prikaz procijenjene logističke funkcije

Testirajmo sada značajnost procijenjenih parametara. Testiramo hipoteze

$$\beta_i = 0,\ i = 0,\ 1. \tag{4.25}$$

Provodimo Waldov test koji je opisan u teorijskom djelu. Testnu statistiku Waldovog testa označavamo s $\mathbf{W}$ ($\mathbf{W} \overset{H_0}{\sim} N(0,1)$). Rezultate testa dajemo u tablici 4.21

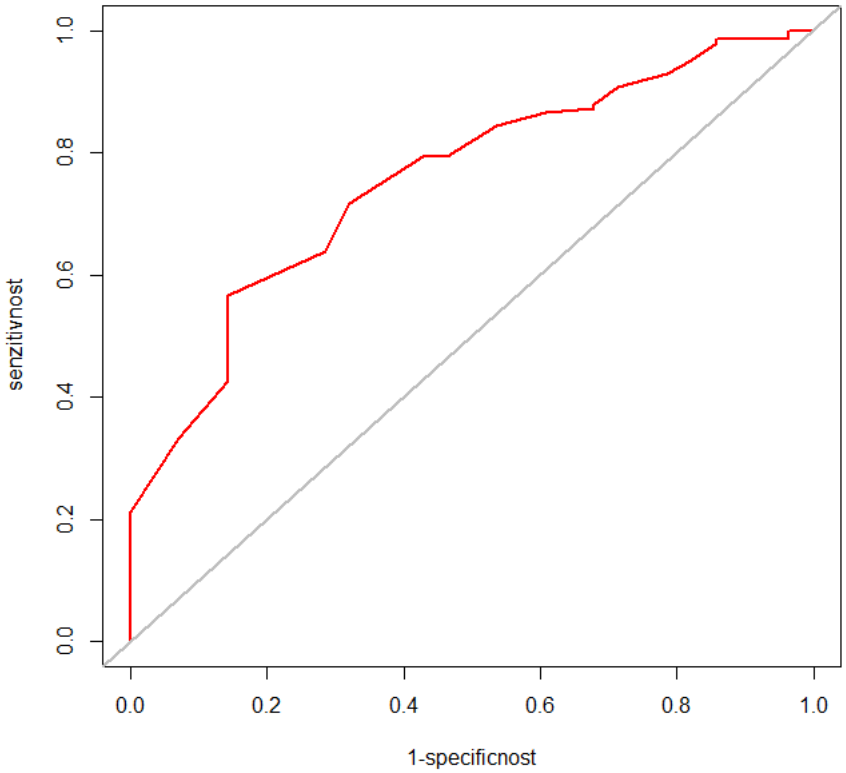
H_0 :	$\mathbf{W}$	$p-$vrijednost
$\beta_0 = 0$	-3.269	0.00108
$\beta_1 = 0$	3.976	0.0316

Tablica 4.21: Testovi značajnosti parametara modela

Iz tablice 4.21 vidimo da hipotezu

$$H_0 : \beta_0 = 0 \tag{4.26}$$

ne možemo odbaciti, dok hipotezu da je $\beta_1 = 0$ odbacujemo na nivou značajnosti $\alpha = 0.05$. Na slici 4.16 dajemo ROC krivulju. Površina ispod ROC krivulje, odnosno AUC promatranog modela iznosi 0.7547.



Slika 4.16: ROC krivulja

4.4 Multivarijatni modeli

Uzorak: 2005. - 2009. godina

Promatramo dio uzorka prije uvođenja državne mature, odnosno promatramo studente koji su upisali studij između 2005. i 2009. godine uključeno. Za svakog studenta promatramo broj ostvarenih bodova u srednjoj školi i broj ostvarenih bodova na prijemnom ispitu kao nezavisne varijable, te uspjeh na prvoj godini studije kao zavisnu, dihotomnu varijablu. Varijablu uspjeh studenta označavamo s $\mathcal{US}$, dok nezavisne varijable označavamo s $\mathcal{BS}$ i $\mathcal{BT}$. Tablica 4.22 daje osnovne podatke o promatranim varijablama, po promatranim godinama.

g	n_g	raspon $\mathcal{BS}$	raspon $\mathcal{BT}$	$\mathcal{US}$
2005	227	0 – 260	0 – 663	0 ili 1
2006	223	0 – 260	0 – 663	0 ili 1
2007	179	0 – 260	0 – 663	0 ili 1
2008	181	0 – 260	0 – 663	0 ili 1
2009	196	0 – 260	0 – 663	0 ili 1
ukupno	1006	0 – 260	0 – 663	0 ili 1

Tablica 4.22: Osnovni podaci za promatrani uzorak

Provodimo logističku regresiju za dani uzorak od $n = 1006$ studenata. Parametre modela procjenjujemo metodom maksimalne vjerodostojnosti. Procijenjeni parametri, kao i procjena standardne pogreške parametara i 95 % pouzdani intervali za procijenjene parametre dani su u tablici 4.23.

	parametar	procjena parametra	$\mathbf{SE}(\hat{\beta}_i)$	95 % interval pouzdanosti
konstanta	$\hat{\beta}_0$	-12.558628656	1.077	$\langle -14.762511901, -10.54033525 \rangle$
$\mathcal{BS}$	$\hat{\beta}_1$	0.039416355	0.004273	$\langle 0.031361798, 0.04811562 \rangle$
$\mathcal{BT}$	$\hat{\beta}_2$	0.007642722	0.0006715	$\langle 0.006368474, 0.00900451 \rangle$

Tablica 4.23: Procjena parametara

Dobivamo da je procijenjena logaritamska transformacija uvjetnog očekivanja oblika

$$\hat{g}(\mathcal{BS}, \mathcal{BT}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathcal{BS} + \hat{\beta}_2 \cdot \mathcal{BT} = -12.559 + 0.0394 \cdot \mathcal{BS} + 0.0076 \cdot \mathcal{BT}, \quad (4.27)$$

odnosno da je procjena uvjetnog očekivanja od dihotomne varijable $\mathcal{US}$ uz dano $\mathcal{BS} = x$ i $\mathcal{BT} = y$ oblika

$$\hat{\pi}(x, y) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y}} = \frac{e^{-12.559 + 0.0394x + 0.0076y}}{1 + e^{-12.559 + 0.0394x + 0.0076y}} \quad (4.28)$$

Testirajmo sada značajnost procijenjenih parametara. Testiramo hipoteze

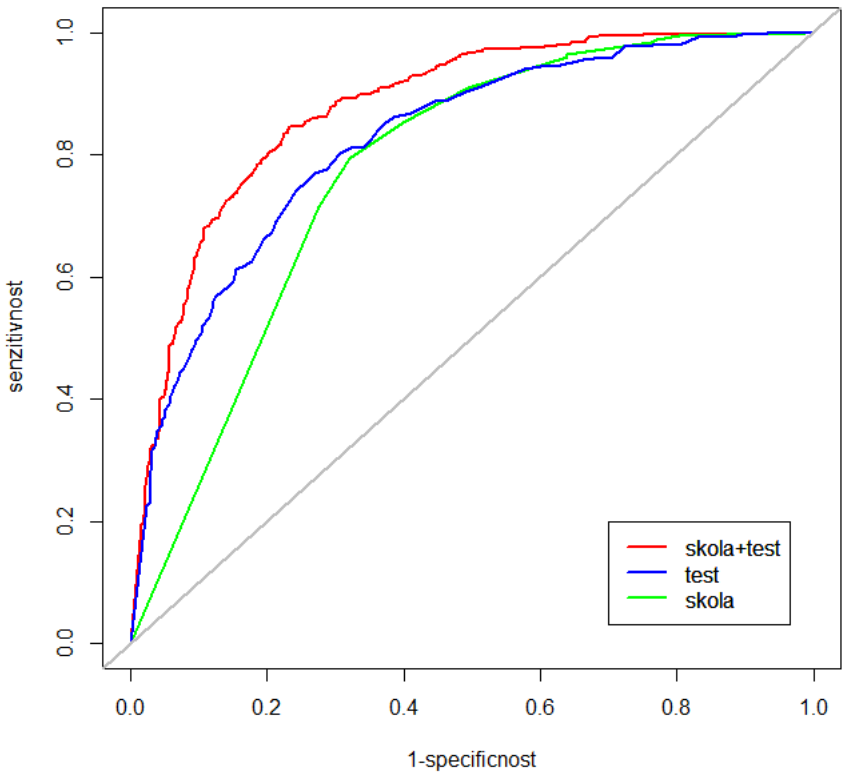
$$\beta_i = 0, \ i = 0, \ 1, \ 2. \quad (4.29)$$

Provodimo Waldov test koji je opisan u teorijskom djelu. Testnu statistiku Waldovog testa označavamo s $\mathbf{W}$ ($\mathbf{W} \overset{H_0}{\sim} N(0, 1)$). Rezultate testa dajemo u tablici 4.24

H_0 :	$\mathbf{W}$	$p-$vrijednost
$\beta_0 = 0$	-11.665	< 0.0001
$\beta_1 = 0$	9.224	< 0.0001
$\beta_2 = 0$	11.381	< 0.0001

Tablica 4.24: Testovi značajnosti parametara modela

Iz tablice 4.24 vidimo da na svakom razumnom novou značajnosti odbacujemo nulte hipoteze, odnosno da su promatrani parametri značajni za model. Na slici 4.17 dajemo ROC krivulju multivarijatnog modela zajedno s pojedinačnim ROC krivuljama univarijatnih modela. Površina ispod ROC krivulje, odnosno AUC promatranog modela iznosi 0.8759.



Slika 4.17: ROC krivulja

Iz grafičkog prikaza 4.17 i usporedbe ROC krivulja vidljivo je da je najbolji model multivarijatni, odnosno onaj koji sadrži i bodove iz škole i bodove prijemnog. Kod univarijatnih modela vidimo da je bolji onaj koji sadrži bodove prijemnog. Zaključujemo da su bodovi prijemnog bolji prediktor uspješnosti studiranja, ali da je optimalno uzeti oba.

Promotrimo sada samo studente koji su uspješno položili sve kolegije prvog semestra studija. Provodimo logističku regresiju za dani uzorak od $n = 484$ studenata. Parametre modela procjenjujemo metodom maksimalne vjerodostojnosti. Procijenjeni parametri, kao i procjena standardne pogreške parametara i 95 % pouzdani intervali za procijenjene parametre dani su u tablici 4.25.

	parametar	procjena parametra	$\mathbf{SE}(\hat{\beta}_i)$	95 % interval pouzdanosti
konstanta	$\hat{\beta}_0$	-7.9775847	1.3391550	$\langle -10.713167155, -5.464378396 \rangle$
$\mathcal{BS}$	$\hat{\beta}_1$	0.0279101	0.0053564	$\langle 0.017812517, 0.038796810 \rangle$
$\mathcal{BT}$	$\hat{\beta}_2$	0.0060854	0.0009735	$\langle 0.004259629, 0.008085831 \rangle$

Tablica 4.25: Procjena parametara

Dobivamo da je procijenjena logaritamska transformacija uvjetnog očekivanja oblika

$$\hat{g}(\mathcal{BS}, \mathcal{BT}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathcal{BS} + \hat{\beta}_2 \cdot \mathcal{BT} = -7.9775847 + 0.0279 \cdot \mathcal{BS} + 0.0061 \cdot \mathcal{BT}, \quad (4.30)$$

odnosno da je procjena uvjetnog očekivanja od dihotomne varijable $\mathcal{US}$ uz dano $\mathcal{BS} = x$ i $\mathcal{BT} = y$ oblika

$$\hat{\pi}(x, y) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y}} = \frac{e^{-7.9775847 + 0.0279x + 0.0061y}}{1 + e^{-7.9775847 + 0.0279x + 0.0061y}} \quad (4.31)$$

Testirajmo sada značajnost procijenjenih parametara. Testiramo hipoteze

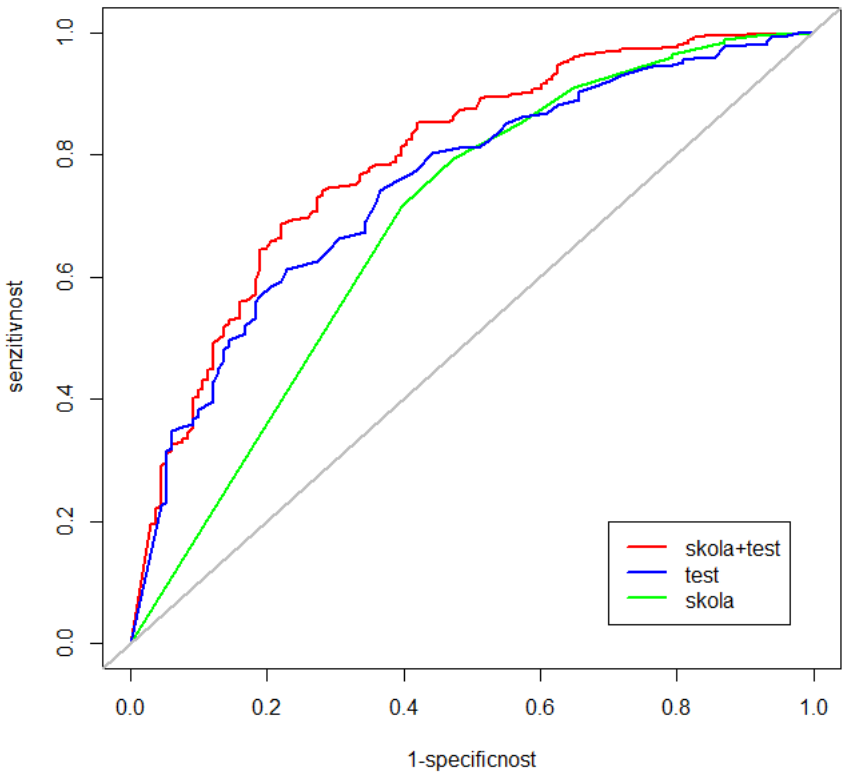
$$\beta_i = 0, \ i = 0, \ 1, \ 2. \quad (4.32)$$

Provodimo Waldov test koji je opisan u teorijskom djelu. Testnu statistiku Waldovog testa označavamo s $\mathbf{W}$ ($\mathbf{W} \overset{H_0}{\sim} N(0, 1)$). Rezultate testa dajemo u tablici 4.26

H_0 :	$\mathbf{W}$	$p-$vrijednost
$\beta_0 = 0$	-5.957	< 0.0001
$\beta_1 = 0$	5.211	< 0.0001
$\beta_2 = 0$	6.251	< 0.0001

Tablica 4.26: Testovi značajnosti parametara modela

Iz tablice 4.26 vidimo da na svakom razumnom nivou značajnosti odbacujemo nulte hipoteze, odnosno da su promatrani parametri značajni za model. Na slici 4.18 dajemo ROC krivulju multivarijatnog modela zajedno s pojedinačnim ROC krivuljama univarijatnih modela. Površina ispod ROC krivulje, odnosno AUC promatranog modela iznosi 0.7904.



Slika 4.18: ROC krivulja

Iz grafičkog prikaza 4.18 i usporedbe ROC krivulja vidljivo je da je najbolji model multivarijatni, odnosno onaj koji sadrži i bodove iz škole i bodove prijemnog. Kod univarijatnih modela vidimo da je bolji onaj koji sadrži bodove prijemnog. Zaključujemo da su bodovi prijemnog bolji prediktor uspješnosti studiranja, ali da je optimalno uzeti oba.

Uzorak: 2010., 2011. godina

Promatramo dio uzorka nakon uvođenja državne mature, odnosno promatramo studente koji su upisali studij 2010. i 2011. godine. Za svakog studenta promatramo broj ostvarenih bodova u srednjoj školi i broj ostvarenih bodova na prijemnom ispitu kao nezavisne varijable, te uspjeh na prvoj godini studije kao zavisnu, dihotomnu varijablu. Varijablu uspjeh studenta označavamo s $\mathcal{US}$, dok nezavisne varijable označavamo s $\mathcal{BS}$ i $\mathcal{BT}$. Tablica 4.27 daje osnovne podatke o promatranim varijablama, po promatranim godinama.

g	n_g	raspon $\mathcal{BS}$	raspon $\mathcal{BT}$	$\mathcal{US}$
2010	198	0 – 300	0 – 600	0 ili 1
2011	197	0 – 300	0 – 600	0 ili 1
ukupno	395	0 – 260	0 – 663	0 ili 1

Tablica 4.27: Osnovni podaci za promatrani uzorak

Provodimo logističku regresiju za dani uzorak od $n = 395$ studenata. Parametre modela procjenjujemo metodom maksimalne vjerodostojnosti. Procijenjeni parametri, kao i procjena standardne pogreške parametara i 95 % pouzdani intervali za procijenjene parametre dani su u tablici 4.28.

	parametar	procjena parametra	$\mathbf{SE}(\hat{\beta}_i)$	95 % interval pouzdanosti
konstanta	$\hat{\beta}_0$	-22.753446	2.495669	$\langle -27.93565231, -18.12991007 \rangle$
$\mathcal{BS}$	$\hat{\beta}_1$	0.033906	0.006607	$\langle 0.02143863, 0.04741172 \rangle$
$\mathcal{BT}$	$\hat{\beta}_2$	0.025474	0.002932	$\langle 0.02002838, 0.03155072 \rangle$

Tablica 4.28: Procjena parametara

Dobivamo da je procijenjena logaritamska transformacija uvjetnog očekivanja oblika

$$\hat{g}(\mathcal{BS}, \mathcal{BT}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathcal{BS} + \hat{\beta}_2 \cdot \mathcal{BT} = -22.753 + 0.0339 \cdot \mathcal{BS} + 0.0255 \cdot \mathcal{BT}, \quad (4.33)$$

odnosno da je procjena uvjetnog očekivanja od dihotomne varijable $\mathcal{US}$ uz dano $\mathcal{BS} = x$ i $\mathcal{BT} = y$ oblika

$$\hat{\pi}(x, y) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y}} = \frac{e^{-22.753 + 0.0339x + 0.0255y}}{1 + e^{-22.753 + 0.0339x + 0.0255y}} \quad (4.34)$$

Testirajmo sada značajnost procijenjenih parametara. Testiramo hipoteze

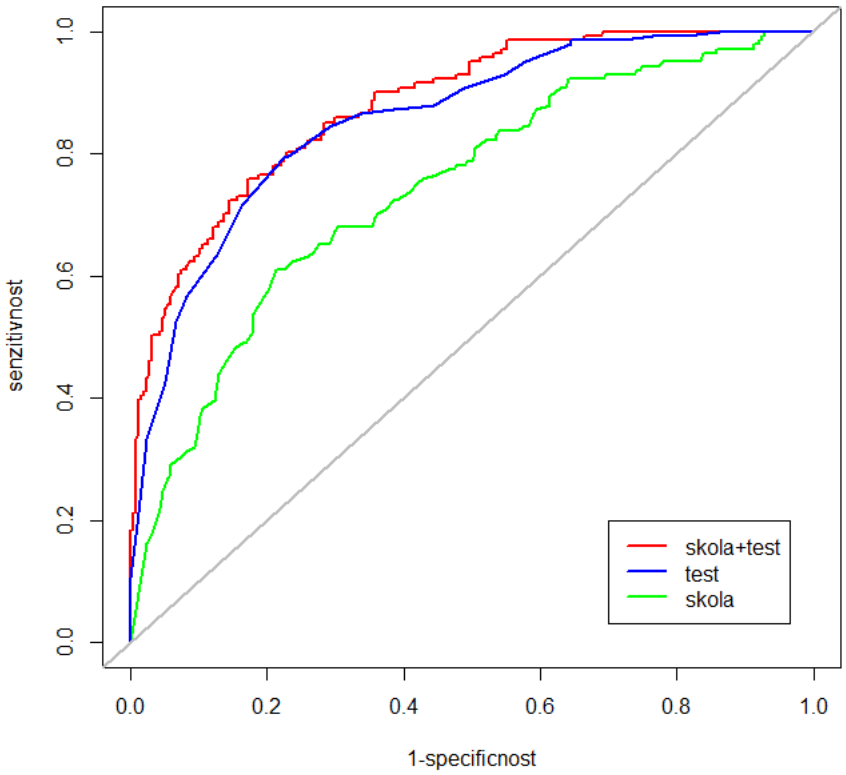
$$\beta_i = 0, \ i = 0, \ 1, \ 2. \quad (4.35)$$

Provodimo Waldov test koji je opisan u teorijskom djelu. Testnu statistiku Waldovog testa označavamo s $\mathbf{W}$ ($\mathbf{W} \overset{H_0}{\sim} N(0, 1)$). Rezultate testa dajemo u tablici 4.29

H_0 :	**W**	$p-$vrijednost
$\beta_0 = 0$	-9.117	< 0.0001
$\beta_1 = 0$	5.132	< 0.0001
$\beta_2 = 0$	8.690	< 0.0001

Tablica 4.29: Testovi značajnosti parametara modela

Iz tablice 4.29 vidimo da na svakom razumnom novou značajnosti odbacujemo nulte hipoteze, odnosno da su promatrani parametri značajni za model. Na slici 4.19 dajemo ROC krivulju. Površina ispod ROC krivulje, odnosno AUC promatranog modela iznosi 0.8786.



Slika 4.19: ROC krivulja

Iz grafičkog prikaza 4.19 i usporedbe ROC krivulja vidljivo je da je najbolji model multivarijatni, odnosno onaj koji sadrži i bodove iz škole i bodove prijemnog. Kod univarijatnih modela vidimo da je bolji onaj koji sadrži bodove prijemnog. Zaključujemo da su bodovi prijemnog bolji prediktor uspješnosti studiranja. Iako je površina ispod ROC krivulje multivarijatnog modela veća nego površina kod uniravijatnog modela s varijablom prijemni, njihova razlika iznosi tek 0.02263081.

Promotrimo sada samo studente koji su uspješno položili sve kolegije prvog semestra studija.

Provodimo logističku regresiju za dani uzorak od $n = 169$ studenata. Parametre modela procjenjujemo metodom maksimalne vjerodostojnosti. Procijenjeni parametri, kao i procjena standardne pogreške parametara i 95 % pouzdani intervali za procijenjene parametre dani su u tablici 4.30.

	parametar	procjena parametra	$\mathbf{SE}(\hat{\beta}_i)$	95 % interval pouzdanosti
konstanta	$\hat{\beta}_0$	-5.813185	1.543291	$\langle -9.022434792, -2.977269393 \rangle$
$\mathcal{BS}$	$\hat{\beta}_1$	0.005178	0.001056	$\langle 0.0083917650.032744069 \rangle$
$\mathcal{BT}$	$\hat{\beta}_2$	0.019879	0.006221	$\langle 0.003199109, 0.007354508 \rangle$

Tablica 4.30: Procjena parametara

Dobivamo da je procijenjena logaritamska transformacija uvjetnog očekivanja oblika

$$\hat{g}(\mathcal{BS}, \mathcal{BT}) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathcal{BS} + \hat{\beta}_2 \cdot \mathcal{BT} = -5.813185 + 0.005178 \cdot \mathcal{BS} + 0.019879 \cdot \mathcal{BT}, \quad (4.36)$$

odnosno da je procjena uvjetnog očekivanja od dihotomne varijable $\mathcal{US}$ uz dano $\mathcal{BS} = x$ i $\mathcal{BT} = y$ oblika

$$\hat{\pi}(x, y) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 y}} = \frac{e^{-5.813185 + 0.005178x + 0.019879y}}{1 + e^{-5.813185 + 0.005178x + 0.019879y}} \quad (4.37)$$

Testirajmo sada značajnost procijenjenih parametara. Testiramo hipoteze

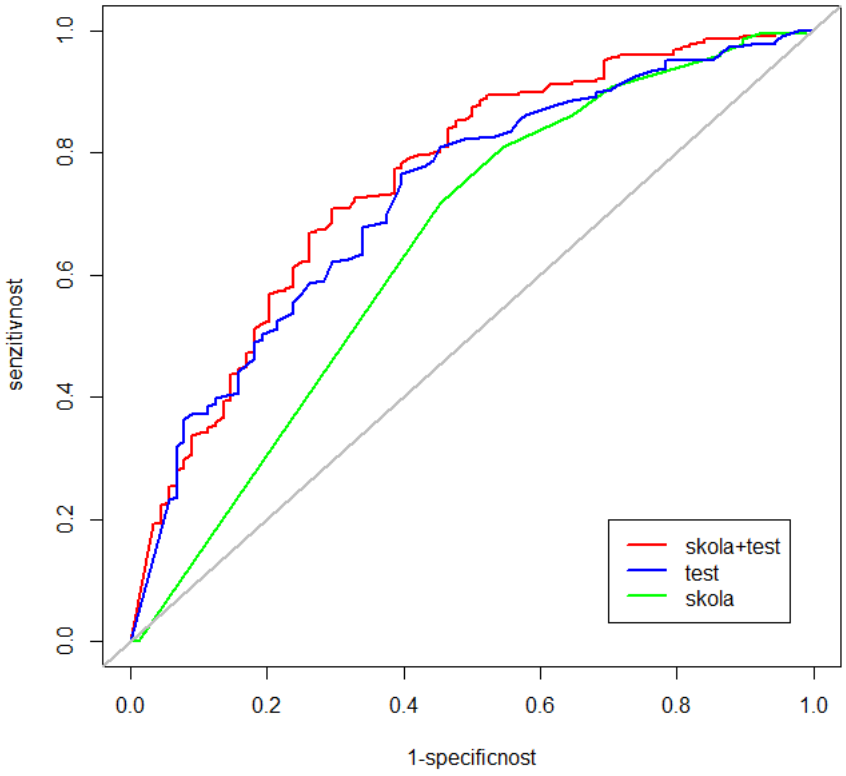
$$\beta_i = 0, \ i = 0, \ 1, \ 2. \quad (4.38)$$

Provodimo Waldov test koji je opisan u teorijskom djelu. Testnu statistiku Waldovog testa označavamo s $\mathbf{W}$ ($\mathbf{W} \overset{H_0}{\sim} N(0, 1)$). Rezultate testa dajemo u tablici 4.31

H_0 :	**W**	$p-$vrijednost
$\beta_0 = 0$	-3.767	0.000165
$\beta_1 = 0$	3.195	< 0.0001
$\beta_2 = 0$	4.905	0.001396

Tablica 4.31: Testovi značajnosti parametara modela

Iz tablice 4.31 vidimo da na svakom razumnom novou značajnosti odbacujemo nulte hipoteze, odnosno da su promatrani parametri značajni za model. Na slici 4.20 dajemo ROC krivulju multivarijatnog modela zajedno s pojedinačnim ROC krivuljama univarijatnih modela. Površina ispod ROC krivulje, odnosno AUC promatranog modela iznosi 0.7627.



Slika 4.20: ROC krivulja

Iz grafičkog prikaza 4.20 i usporedbe ROC krivulja vidljivo je da je najbolji model multivarijatni, odnosno onaj koji sadrži i bodove iz škole i bodove prijemnog. Kod univarijatnih modela vidimo da je bolji onaj koji sadrži bodove prijemnog. Zaključujemo da su bodovi prijemnog bolji prediktor uspješnosti studiranja, ali da je optimalno uzeti oba.

Poglavlje 5

Dodatak: Procjena parametara modela logističke regresije

Uvod i opće oznake

Varijablu odaziva Y zovemo dihotomna ili binarna ukoliko poprima samo dva moguća stanja. Takve varijable u pravilu kodiramo s dva stanja: $y = 0$ ili $y = 1$. Metode logističke regresije opisuju odnos između dihotomne varijable odaziva (Y) i jedne ili više varijabli poticaja ($X_1, X_2, \ldots, X_k, k \in \mathbb{N}$). Uvjetno očekivanje varijable odaziva Y uz danu vrijednost varijable poticaja $X = x$ označavamo s $\mathbb{E}(Y|X = x)$.
Ukoliko je Y dihotomna varijabla, raspisivanjem uvjetnog očekivanja dobivamo

$$\mathbb{E}(Y|X = x) = 0 \cdot \mathbb{P}(Y = 0|X = x) + 1 \cdot \mathbb{P}(Y = 1|X = x) = \mathbb{P}(Y = 1|X = x), \tag{5.1}$$

što nam daje interpretaciju uvjetnog očekivanja zavisne varijable Y uz dano $X = x$ kao uvjetne vjerojatnosti da zavisna varijabla Y poprimi vrijednost 1 uz dano $X = x$.
Slijedi da dihotomna varijabla odaziva $Y|X = x$ ima Bernoullijevu radiobu, odnosno da vrijedi

$$Y|X = x \sim \begin{pmatrix} 0 & 1 \\ 1 - \mathbb{E}(Y|X = x) & \mathbb{E}(Y|X = x) \end{pmatrix}. \tag{5.2}$$

U slučaju kada je Y dihotomna varijabla $\mathbb{E}(Y|X = x)$ nalazi se u segmentu $[0, 1]$, pa se $\mathbb{E}(Y|X = x)$ postepeno približava 0 (s desna), odnosno 1 (s lijeva).

Kako bi pojednostavili notaciju, uvodimo oznaku $\pi(x) = \mathbb{E}(Y|X = x)$ za uvjetno očekivanje od Y uz dano $X = x$ kada koristimo logističku distribuciju.

Univarijatni model (k=1)

Pomoću logističke funkcije (s određenom supstitucijom) dobivamo općenitu formulu za uvjetno očekivanje od Y uz dano $X = x$, odnosno vrijedi

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{5.3}$$

Nadalje, uvodimo transformaciju uvjetnog očekivanja $\pi(x)$ u oznaci $g(x) := G(\pi(x))$, gdje je $G : \mathbb{R} \to \mathbb{R}$ definirana kao

$$G(x) = \ln\left(\frac{x}{1 - x}\right). \tag{5.4}$$

Preslikavanje g nazivamo logaritamska transformacija (logit transformation) uvjetnog očekivanja $\pi(x)$ i dobivamo

$$g(x) = G(\pi(x)) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \ln\left(\frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{1 - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}\right) = \beta_0 + \beta_1 x \tag{5.5}$$

Neka je X nezavisna varijabla, Y zavisna dihotomna varijabla i neka je dani uzorak on $n \in \mathbb{N}$ nezavisnih opservacija (x_i, y_i), $i \in \{1, 2, \ldots, n\}$. Iz relacije 5.5 slijedi kako je logaritamska transformacija uvjetnog očekivanja $\pi(x)$ oblika

$$g(x) = \beta_0 + \beta_1 x. \tag{5.6}$$

Želimo na temelju danog uzorka (x_i, y_i), $i \in \{1, 2, \ldots, n\}$ procijeniti vektor parametara $\boldsymbol{\beta} := (\beta_0, \beta_1)$.

U linearnim regresijskim modelima najbolji, linearni, nepristrani procjenitelj (BLUE) za $\boldsymbol{\beta} := (\beta_0, \beta_1)$ je LS-procjenitelj $\hat{\boldsymbol{\beta}} = (\hat{\beta_0}, \hat{\beta_1})$, odnosno procjenitelj dobiven metodom najmanjih kvadrata (least squares method).

Kako u modelu s dihotomnom varijablom odaziva Y neće vrijediti Gauss-Markovljevi uvjete, ne možemo primijeniti metodu najmanjih kvadrata.

Općenita metoda procjene za vektor parametara $\boldsymbol{\beta}$ je metoda maksimalne vjerodostojnosti.

Pod pretpostavkom da su opažanja nezavisna definiramo funkciju vjerodostojnosti, u oznaci ℓ, kao

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n} \xi(x_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x))^{1-y_i}. \tag{5.7}$$

Nadalje, za funkciju vjerodostojnosti ℓ definiramo funkciju log-vjerodostojnosti, u oznaci $\mathcal{L}$, kao

$$\mathcal{L}(\boldsymbol{\beta}) = \ln\left(\ell(\boldsymbol{\beta})\right) = \sum_{i=1}^{n} \Big(y_i \ln\left(\pi(x_i)\right) + (1 - y_i) \ln\left(1 - \pi(x_i)\right)\Big) \tag{5.8}$$

Princip maksimalne vjerodostojnosti procjenjuje vektor parametara $\boldsymbol{\beta}$ s vektorom parametara $\hat{\boldsymbol{\beta}}$ koji maksimizira funkciju vjerodostojnosti, ondosno s $\hat{\boldsymbol{\beta}}$ za koji vrijedi

$$\ell(\hat{\boldsymbol{\beta}}) = \max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}). \tag{5.9}$$

Kako je matematički jednostavnije maksimizirati funkciju log-vjerodostojnosti tražimo $\hat{\boldsymbol{\beta}}$ za koji vrijedi

$$\mathcal{L}(\hat{\boldsymbol{\beta}}) = \max_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}). \tag{5.10}$$

Parcijalno defiriviramo funkciju log-vjerodostojnosti $\mathcal{L}$ u odnosu na β_0 i β_1 i izjednačavamo s nulom. Jednadžbe koje dobivamo zovemo jedndžbe vjerodostojnosti i one su

$$\frac{\partial}{\partial \beta_0} \mathcal{L}(\beta_0, \beta_1) = 0 \iff \sum_{i=1}^{n} (y_i - \pi(x_i)) = 0 \tag{5.11}$$

i

$$\frac{\partial}{\partial \beta_1}\mathcal{L}(\beta_0, \beta_1) = 0 \iff \sum_{i=1}^{n} x_i(y_i - \pi(x_i)) = 0 \tag{5.12}$$

Jednadžbe vjerodostojnosti 5.11 i 5.12 su nelinearne u parametrima β_0 i β_1. Jedna od metode za rješavanje tih jednadžbi je iterativna težinska metoda najmanjih kvadrata (eng. iteratively reweighted least squares; IRLS).

Pretvorimo li x_i u vektor $\tilde{x}_i = \begin{bmatrix} 1 & x_i \end{bmatrix}^T$, $i = \{1, \ldots, n\}$ uz navedene oznake dobivamo

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \; X = \begin{bmatrix} \tilde{x}_1^T \\ \vdots \\ \tilde{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \; Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \tag{5.13}$$

pa po pretpostavci modela logističke regresije imamo

$$\pi(\tilde{x}_i; \boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}^T \tilde{x}_i}}{1 + e^{\boldsymbol{\beta}^T \tilde{x}_i}}, \; 1 - \pi(\tilde{x}_i; \boldsymbol{\beta}) = \frac{1}{1 + e^{\boldsymbol{\beta}^T \tilde{x}_i}}. \tag{5.14}$$

Dobivamo da je

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big(y_i \boldsymbol{\beta}^T x_i - \ln(1 + e^{\boldsymbol{\beta}^T \tilde{x}_i}) \Big), \tag{5.15}$$

odnosno da su jednadžbe vjerodostojnosti oblika

$$\frac{\partial}{\partial \beta_j}\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} x_{i,j+1}(y_i - \pi(\tilde{x}_i; \boldsymbol{\beta})) = 0, \; j = 0, 1, \tag{5.16}$$

što matrično zapisujemo kao

$$\frac{\partial}{\partial \boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \tilde{x}_i(y_i - \pi(\tilde{x}_i; \boldsymbol{\beta})) = 0 \tag{5.17}$$

Sustav od $k + 1 = 2$ jednadžbi rješavamo Newton-Raphsonovom metodom. Kako je za $k, j \in \{0, 1\}$

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} x_{i,j+1} x_{i,k+1} \pi(\tilde{x}_i; \boldsymbol{\beta})(1 - \pi(\tilde{x}_i; \boldsymbol{\beta})), \tag{5.18}$$

dobivamo da je Hesseova matrica, u oznaci $\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$ oblika

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^T \pi(\tilde{x}_i; \boldsymbol{\beta})(1 - \pi(\tilde{x}_i; \boldsymbol{\beta})). \tag{5.19}$$

Za dani β^{old} jedna Newton-Raphsonova iteracija je

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)^{-1} \frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \tag{5.20}$$

Neka je $\mathbb{X}$ ulazna matrica (dimenzija $n \cdot (k+1)$) i $\mathbb{Y}$ vektor varijabli odaziva.
Neka je $\mathbf{p}$ vektor uvjetnih očekivanja uz dani parametar β^{old}, te $\mathbb{W}$ dijagonalna matrica težina (dimenzija $n \cdot n$) tako da je $i-$ti dijagonalni element oblika $\pi(\tilde{x}_i;\beta^{old})(1-\pi(\tilde{x}_i;\beta^{old}))$, odnosno

$$\mathbf{p} = \begin{bmatrix} \pi(\tilde{x}_1;\beta^{old}) \\ \vdots \\ \pi(\tilde{x}_n;\beta^{old}) \end{bmatrix}, \ \mathbb{W} = \begin{bmatrix} \pi(\tilde{x}_1;\beta^{old})(1-\pi(\tilde{x}_1;\beta^{old})) & 0 & \dots & 0 \\ 0 & \ddots & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & \pi(\tilde{x}_n;\beta^{old})(1-\pi(\tilde{x}_n;\beta^{old})) \end{bmatrix}. \tag{5.21}$$

Tada matrice parcijalnih derivacija možemo zapisati matrično kao

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbb{X}^T(\mathbb{Y} - \mathbf{p}) \tag{5.22}$$

i

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\mathbb{X}^T \mathbb{W} \mathbb{X}. \tag{5.23}$$

Dobivamo da je jedna Newton-Raphsonova iteracija oblika

$$\beta^{new} = \beta^{old} + (\mathbb{X}^T \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^T (\mathbb{Y} - \mathbf{p}) = (\mathbb{X}^T \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^T \mathbb{W} \mathbf{z}, \tag{5.24}$$

gdje je $\mathbf{z} = (\mathbb{X}\beta^{old} + \mathbb{W}^{-1}(\mathbb{Y} - \mathbf{p}))$. Ukoliko se $\mathbf{z}$ može napisati kao varijabla odaziva matrice $\mathbb{X}$, dobivamo da je β^{new} rješenje problema težinskih namjanjih kvadrata:

$$\beta^{new} = \min_{\beta} (\mathbf{z} - \mathbb{X}\beta)^T \mathbb{W} (\mathbf{z} - \mathbb{X}\beta) \tag{5.25}$$

Multivarijatni model

Neka je $k \in \mathbb{N}$. Neka su za $i = 1, \dots, n$ dani vektori poticaja (s dodanom jedinicom) X_i, $\boldsymbol{\beta}$ vektor parametara i $\mathbb{Y}$ vektor odaziva. Matrično imamo

$$\mathbb{X} = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}, \ \mathbb{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}. \tag{5.26}$$

Kako smo univarijatni model radili u matričnom obliku, multivarijatni model slijedi analogno. U nastavku raspisujemo općeniti algoritam za multivarijatni model.

Općeniti algoritam

i) stavimo $\beta = 0$;

ii) izračunamo elemente vektora $\mathbf{p}$, gdje je $i-$ti element oblika

$$\pi(X_i;\beta) = \frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}} i = 1, \ldots, n, \tag{5.27}$$

gdje je $X_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix}$ $i-$ti redak matrice $\mathbb{X}$;

iii) izračunamo dijagonalne elemente dijagonalne matrice $\mathbb{W}$, gdje je $i-$ti dijagonalni element oblika

$$\pi(X_i;\beta)(1 - \pi(X_i;\beta)), i = 1, \ldots, n; \tag{5.28}$$

iv) $\mathbf{z} = \mathbb{X}\beta + (\mathbb{W})^{-1}(\mathbb{Y} - \mathbf{p})$;

v) $\beta = (\mathbb{X}^T\mathbb{W}\mathbb{X})^{-1}\mathbb{X}^T\mathbb{W}\mathbf{z}$;

vi) ukoliko je kriterij zaustavljanja zadovoljen, završavamo; u suprotnom, vraćamo se na korak ii);

Kako je $\mathbb{W}$ dijagonalna matrica reda n, računanje s njom je neefikasno. Uvodimo matricu $\tilde{\mathbb{X}}$ kao

$$\tilde{\mathbb{X}} = \begin{bmatrix} \pi(X_1;\beta)(1 - \pi(X_1;\beta))X_1^T \\ \pi(X_2;\beta)(1 - \pi(X_2;\beta))X_2^T \\ \vdots \\ \pi(X_n;\beta)(1 - \pi(X_n;\beta))X_n^T \end{bmatrix}, \tag{5.29}$$

te dobivamo efikasniji algoritam:

i) stavimo $\beta = 0$;

ii) izračunamo elemente vektora $\mathbf{p}$, gdje je $i-$ti element oblika

$$\pi(X_i;\beta) = \frac{e^{\beta^T X_i}}{1 + e^{\beta^T X_i}} i = 1, \ldots, n, \tag{5.30}$$

gdje je $X_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix}$ $i-$ti redak matrice $\mathbb{X}$;

iii) izračunamo matricu $\tilde{\mathbb{X}}$;

iv) $\beta = \beta + (\mathbb{X}^T\tilde{\mathbb{X}})^{-1}\mathbb{X}^T(\mathbb{Y} - \mathbf{p})$;

v) ukoliko je kriterij zaustavljanja zadovoljen, završavamo; u suprotnom, vraćamo se na korak ii);

Implementacija koda u Matlab-u

Funkcija (nazvana **IRLS**) u Matlab-u prima matricu poticaja $\mathbb{X}$ (svaki redak je jedno opažanje, stupci su varijable poticaja) dimenzija $n \cdot k$ gdje je n broj opažanja, a k broj varijabli poticaja, prirodni broj m (željeni broj iteracija), te dihotomni vektor odaziva Y.
IRLS vraća vektor parametara $\boldsymbol{\beta}$, te crta graf konvergencije svakog parametra. Početno je vektor $\boldsymbol{\beta}$ zadan kao nul-vektor.

```
function [beta] = IRLS(X, Y, m)
    n=length(X);
    k=size(X);
    k=k(2);
    X_novi=[ones(n,1) X];
    beta=zeros(k+1,1);
    pi=zeros(n,1);
    hold on;
    myColor=rand(k+1,3);

    for i = 1:m
        pi=exp(X_novi*beta) ./ (1+exp(X_novi*beta));
        pi_novi=pi;
        for j=1:k
           pi_novi=[pi_novi pi];
        end
        X_tilda = pi_novi.*X_novi;
        beta=beta+inv(transpose(X_novi)*X_tilda)*transpose(X_novi)*(Y-pi);
        for j=1:k+1
            plot(i,beta(j),'*','Color',myColor(j,:))
            xlabel('broj iteracija')
            ylim([-50, 50])
        end
    end
```

```
    for j=1:k+1
        text(m-10*j,beta(j)+5, ['beta' num2str(j)])
    end
end
```

Bibliografija

[1] D. W. Hosmer, JR & S. Lemeshow & R. X. Sturdivant, *Applied logistic regression*, John Wiley & Sons, Inc., Hobuken, New Jersey, 3rd edition, 2013.

[2] D. D. Wackerly & W. Mendenhall III & R. L. Scheaffer, *Mathematical Statistics with Applications*, Thomson Learning, CA, USA, 7th edition, 2008.

[3] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 2002.

[4] G. K. Bhattacharyya & R. A. Johnson, *Statistical Concepts and Methods*, John Wiley & Sons, Inc., New York, 1977.

Sažetak

U ovom diplomskom radu promatramo uzorak studenata koji su upisali studij matematike u razdoblju od 2005. do 2011. godine. Uzorak razdvajamo u dva djela, prvi prije uvođenja državne mature i drugi za vrijeme. U prvom uzorku postoji promjena u računarskim kolegijima, dok drugi uzorak ima jednake kolegije kroz godine. Definirali smo pojam uspješnog studenta kao studenta koji je položio sve kolegije prve godine u roku. Rad se sastoji od pet poglavlja.

U prvom poglavlju dajemo deskriptivnu statistiku promatranog uzorka. Zaključili smo da ne postoji statistički značajna razlika između generacija koje su slušale različite računarske kolegije (na nivou značajnosti $\alpha = 0.05$), te pretpostavili da su u tom uzorku kolegiji jednaki kroz godine. Nadalje, analizirali smo prvo bodove iz srednje škole (oznaka $\mathcal{BS}$), a zatim i bodove s prijemnog ispita/državne mature (oznaka $\mathcal{BT}$). Uz stupčaste dijagrame, box-plotove i tablice s osnovnim statističkim veličinama, proveli smo test analize varijance (ANOVA). U slučaju kada je postojala statistički značajna razlika (za nivo značajnosti $\alpha = 0.05$) između barem dvije grupe, proveli smo dodatno Tukey test i rezultate dali pregledno u tablici. Na kraju prvog poglavlja provodimo analizu uspješnih studenata. Promatramo udio uspješnih studenata kroz godine, te položenosti pojedinih kolegija. Uz grafičke prikaze i osnovne tablice, provodimo testove o proporcijama. Svi rezultati dani su pregledno u tablicama.

Drugo i treće poglavlje daju teorijski aspekt logističke regresije. Dan je uvod u univarijatnu i multivarijatnu analizu, definirani su glavni pojmovi i uvedene opće oznake. Također je dana pozadina prilagodbe modela logističke regresije i testiranja značajnosti koeficijenata. Dana su tri glavna testa: test omjera vjerodostojnosti, Waldov test i Score test. Na kraju poglavlja dana je pozadina procjene intervala pouzdanosti.

Četvrto poglavlje daje primjenu osnovnih univarijatnih i multivarijatnih modela logističke regresije na promatranom uzorku studenata. Promatrali smo utjecaj nezavisnih varijabli $\mathcal{BS}$ i $\mathcal{BT}$ na zavisnu dihotomnu varijablu uspjeha ($\mathcal{US}$). Zaključili smo da postoji veza između uspjeha studenta i promatranih nezavisnih varijabli. Pomoću ROC krivulja

zaključili smo da je nezavisna varijabla $\mathcal{BT}$ bolja u procjeni uspjeha od nezavisne varijable $\mathcal{BS}$, ali da je optimalno promatrati multivarijatni model koji sadrži obje varijable. Uz standardne testove i tablice, dani su grafički prikazi logističkih funkcija zajedno s 95% prugama pouzdanosti i pripadne ROC krivulje. Proveli smo pripadni Waldov test za značajnost koeficijenata i zaključili da su u svim univarijatnim i multivarijatnim modelima promatrane nezavisne varijable značajne. Također smo promatrali i dio uzorka studenata koji su položili sve kolegije prvog semestra studija. U svim modelima varijable su bile značajne (nivo značajnosti $\alpha = 0.05$).

U zadnjem poglavlju dajemo općeniti algoritam za procjenu parametara multivarijatnog modela. Također je dana i implementacija koda u programskom jeziku Matlab.

Summary

In this thesis we look at a sample of students who enrolled in the study of mathematics in the period from 2005 to 2011. The sample is split into two parts, the first before the introduction of the state graduation and the other after. In the first sample there is a change in computer courses, while the second sample has the same courses over the years. We have defined the concept of a successful student as a student who has passed all courses in the first year of the term. Thesis consists of five chapters.

In the first chapter we provide descriptive statistics of the sample. We concluded that there was no statistically significant difference (significance level is $\alpha = 0.05$) between the generation that listened to various computer courses and we assumed that in the sample courses are the same through the years. Furthermore, we analyzed the first varijable, points from high school ($\mathcal{BS}$), and then points to the entrance examination/State graduation ($\mathcal{BT}$). With bar charts, box-plots and tables with basic properties, we conducted a test of analysis of variance (ANOVA). In the case where there was a statistically significant difference (significance level is $\alpha = 0.05$) between at least two groups, we conducted further Tukey test and the results are given in the tables. At the end of the first chapter we conduct an analysis of successful students. We looked at the share of successful students over the years and at individual courses. With graphics and basic tables, we conducted tests of proportions. All results are presented in tables.

The second and third chapter gives the theoretical aspect of the logistic regression. An introduction to univariate and multivariate analysis is givena and main concepts are defined. There is also a background of adapting logistic regression model and testing the significance of the coefficients with three tests: likelihood ratio test, Wald test and Score test. At the end of the chapter we give background of estimation of confidence intervals.

The fourth chapter provides the application of basic univariate and multivariate logistic regression model to the observed sample of students. We looked at the impact of the independent variables $\mathcal{BS}$ and $\mathcal{BT}$ on the dependent dichotomous variable success ($\mathcal{US}$). We concluded that there is a link between student success and the observed independent

variables. Using the ROC curve, we have concluded that the independent variable $\mathcal{BT}$ is better in assessing the success of the independent variables $\mathcal{BS}$, but that it is optimally to use multivariate model that includes both variables. In addition to the standard tests and tables, we give graphical representations of logistics functions with 95% stripes reliability and associated ROC curve. We conducted the corresponding Wald test for the significance of the coefficients and concluded that in all univariate and multivariate models observed independent variables are significant. We also looked at part of the sample of students who have passed all courses in the first semester. All model variables were significant (significance level $\alpha = 0.05$).

In the last chapter we give a general algorithm for estimating parameters of multivariate models. There is also an implementation of the code in the programming language Matlab.

Životopis

Zovem se Mislav Zorko. Rođen sam u Koprivnici 21. lipnja 1990. godine. Djetinjstvo sam proveo u Križevcima, gradu 57 kilometara udaljenom od Zagreba. Osnovnoškolsko obrazovanje završio sam 2005. godine u Osnovnoj školi Ljudevita Modeca Križevci. Srednjoškolsko obrazovanje završio sam 2009. godine u Gimnaziji Ivana Zakmardija Dijankovečkoga Križevci. Preddiplomski sveučilišni studij matematike završio sam 2013. godine, te upisao Diplomski sveučilišni studij Matematičke statistike.